RECONSTRUCTION AS SPATIAL EVALUATION: DIAGNOSING CROSS-VIEW SPATIAL INTELLIGENCE IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial intelligence fundamentally requires a model to combine local visual observations into a geometrically consistent representation of the physical world. Many static benchmarks assess isolated spatial judgments, while interactive embodied evaluations combine perception with exploration, planning, and control. To bridge this gap, we introduce **RECONSPATIAL**, a benchmark centered on multi-view 3D scene reconstruction across indoor, tabletop, and outdoor environments. Given real multi-view images and target-object queries, large vision-language models (LVLMs) jointly predict oriented 3D bounding boxes in a shared, camera-anchored metric coordinate frame. Reconstructions are evaluated along two complementary axes: *Structure* for relative layout and *Metric* for physical scale and camera-anchored placement. Nine atomic spatial QA tasks, grouped into *Geometric Perception* and *Spatial Relations*, probe spatial capabilities involved in joint reconstruction. Offline navigation extends evaluation to goal-directed planning and provides a setting to examine reconstructed maps as intermediate reasoning representations. Evaluation of 18 frontier LVLMs reveals a consistent gap between relative scene structure and metric reconstruction; even GPT-6-Astra scores 81.4% on Structure but only 58.7% on Metric. Diagnostic analyses link atomic spatial capabilities to reconstruction quality, while showing that relational consistency alone does not ensure metric accuracy. Explicit 3D maps provide an intermediate representation that separates scene construction from path planning, revealing how geometric errors limit navigation success. Finally, spatial fine-tuning yields broad transfer across both spatial and general vision-language benchmarks.

1 INTRODUCTION

A central requirement of spatial intelligence is the ability to combine partial, viewpoint-dependent observations into a coherent representation of three-dimensional space. Although contemporary Large Vision-Language Models (LVLMs) have made substantial progress in visual-semantic understanding (Bai et al., 2025; Comanici et al., 2025; Achiam et al., 2023; Fu et al., 2025; 2026), their spatial intelligence remains severely limited, particularly in understanding complex 3D environments (Yang et al., 2025; 2026a).

Contemporary spatial benchmarks primarily fall into two main paradigms. Static benchmarks isolate specific geometric attributes—such as relative direction, depth, or viewpoint transformation (Yang et al., 2025; 2026a; Wang et al., 2025; Jia et al., 2026; Li et al., 2025a). However, independent queries fail to test whether a model can jointly recover a coherent scene: for instance, knowing which object is closer does not determine whether their 3D positions and extents form a valid layout. Conversely, embodied benchmarks evaluate spatial understanding through interaction in simulated environments (Majumdar et al., 2024; Hong et al., 2026), where outcomes jointly depend on perception, exploration, planning, control, and simulator-specific execution. This leaves an unmet need: evaluating whether models can consolidate cross-view evidence into a geometrically consistent 3D scene, independent of interactive execution.

To bridge this evaluation gap, we introduce RECONSPATIAL, which establishes *multi-view 3D scene reconstruction* as its core evaluation task. Given multi-view observations and target object queries, an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

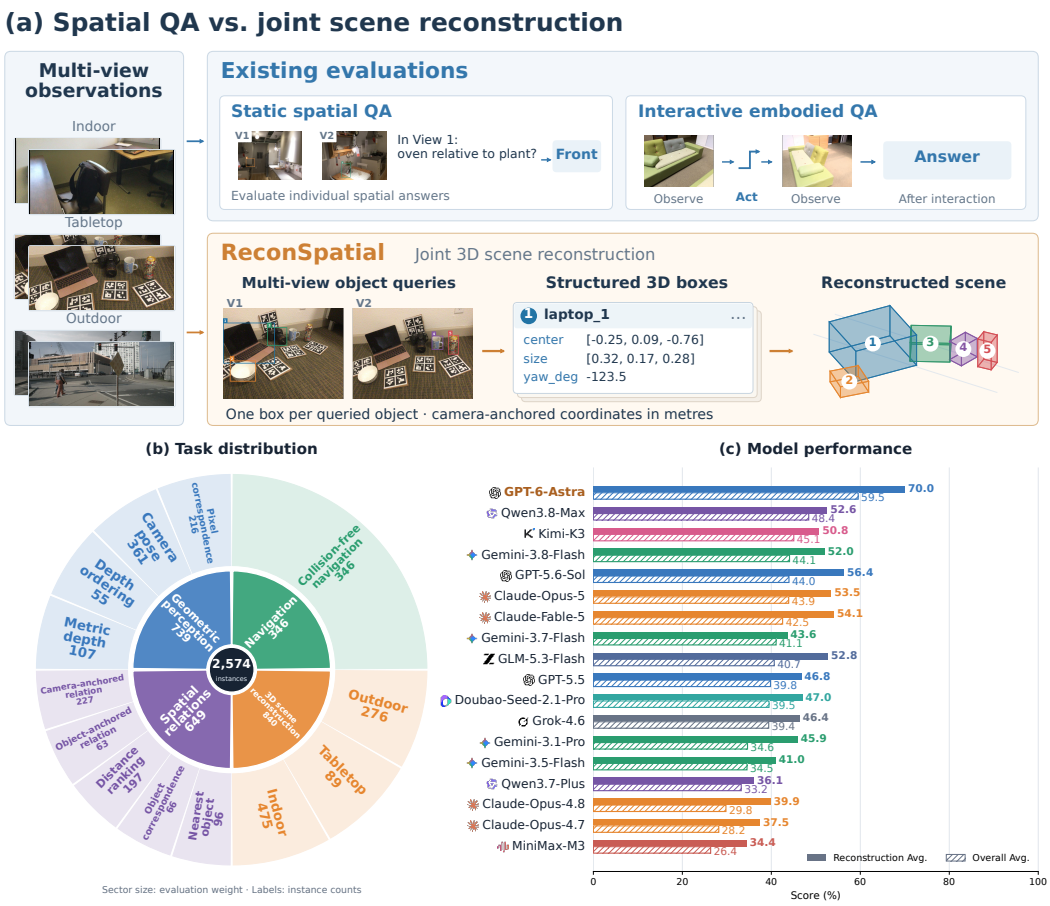


Figure 1: From spatial QA to joint scene reconstruction. **(a)** Static and interactive embodied QA compared with reconstruction from multi-view object queries via structured 3D boxes, visualized as a joint scene; reference geometry illustrates the target output. **(b)** Task distribution across four evaluation levels. **(c)** Eighteen models ranked by Overall, alongside their Reconstruction average.

LVLM is tasked with predicting oriented 3D bounding boxes in a shared, camera-anchored metric coordinate frame. This joint objective requires models to associate objects across viewpoints, resolve relative camera geometry, and estimate metric extents simultaneously. Providing queried 2D regions controls for open-set detection, isolating spatial reasoning as the primary point of evaluation. We assess each reconstruction along two complementary axes: *Structure*, which evaluates relative layout and topological arrangement, and *Metric*, which measures physical scale and camera-anchored 3D placement across indoor, tabletop, and outdoor scenes.

Around this core reconstruction task, RECONSPATIAL incorporates two complementary evaluation suites. Nine spatial QA tasks—grouped into *Geometric Perception* and *Spatial Relations*—isolate constituent capabilities such as cross-view correspondence, camera pose, and relative depth. Their scores help us examine how individual spatial capabilities relate to reconstruction performance. In parallel, an offline *Navigation* task provides an independent evaluation of spatial planning: given multi-view observations, a starting viewpoint, and a visual target, the model predicts a path that reaches the target while avoiding obstacles. Together, these tasks assess local spatial judgments, joint scene reconstruction, and goal-directed planning (Figure 1).

Evaluation of 18 frontier LVLMs reveals a consistent gap between relative scene structure and metric reconstruction; even the top-performing GPT-6-Astra scores 81.4% on Structure but only 58.7% on Metric. Diagnostic analyses connect individual spatial capabilities to reconstruction quality, showing that relational consistency alone does not ensure metric accuracy. Furthermore, explicit 3D maps separate scene construction from path planning: although self-generated maps

aid navigation, replacing them with ground-truth geometry boosts success rates by an additional 55.2 to 69.7 percentage points. This gap indicates that downstream navigation is primarily bounded by perceptual accuracy rather than planning capability. Finally, spatial fine-tuning confers broad performance gains across both spatial and general vision-language benchmarks.

The contributions of this work are summarized as follows:

- **Joint 3D reconstruction for spatial evaluation.** We introduce RECONSPATIAL, which establishes a new evaluation paradigm that benchmarks LVLMs on jointly recovering object-level 3D bounding boxes from real multi-view observations. Complementary *Structure* and *Metric* scores distinguish relative scene coherence from physical scale and placement.
- **A hierarchical diagnostic suite across diverse domains.** Covering indoor, tabletop, and outdoor scenes, RECONSPATIAL unifies 1,388 atomic items, 840 reconstruction instances, and 346 offline navigation episodes to evaluate local spatial capabilities, joint scene reconstruction, and spatial planning.
- **Diagnostic Findings and Downstream Decoupling.** Evaluating eighteen frontier models uncovers a pervasive Structure–Metric disparity. Controlled interventions isolate metric scale as the core bottleneck, downstream map substitution decouples spatial perception from planning competence, and spatial fine-tuning confirms transferable reasoning benefits.

2 RELATED WORK

Benchmarking Spatial Reasoning. Spatial reasoning benchmarks have primarily targeted local geometric queries (Yang et al., 2025; Ma et al., 2025; Wang et al., 2025; Jia et al., 2026) or embodied task execution (Majumdar et al., 2024; Hong et al., 2026; Zhao et al., 2026b). Recent efforts also explore single-image inverse rendering (Liu et al., 2025a), 3D room generation (Yang et al., 2026b), and floor-plan or metric-map reconstruction (Petersson et al., 2025; Huang et al., 2025). In contrast, RECONSPATIAL evaluates joint object geometry from real-world multi-view observations, directly scoring predicted 3D bounding boxes against annotated ground truth. Complementary *Structure* and *Metric* scores explicitly disentangle relative scene layout from physical scale and camera-anchored placement.

Spatial Representations for Reasoning and Planning. Explicit spatial representations provide essential intermediate abstractions between perception and action. Beyond end-to-end latent reasoning, recent approaches interleave symbolic reasoning with visual sketching (Wu et al., 2025a), while cognitive-mapping approaches study structured representations that scaffold mental navigation (Yang et al., 2025; Zhu et al., 2026). Complementary work examines data and training strategies for spatial reasoning (Ouyang et al., 2025; Zhao et al., 2026a). We build upon this insight by treating reconstructed 3D bounding box maps as an explicit spatial reasoning representation, directly probing how their geometric fidelity scaffolds downstream planning.

3 RECONSTRUCTION AS SPATIAL EVALUATION

RECONSPATIAL centers on joint 3D scene reconstruction from multiple views. Spatial QA tasks provide complementary measurements of individual spatial skills, while an independent navigation task evaluates goal-directed planning (Figure 2).

3.1 PROBLEM FORMULATION

Let $\mathcal{V} = (I_1, \ldots, I_N)$ denote an ordered sequence of multi-view observations, with I_1 serving as the reference view. Let $\mathcal{Q} = \{q_k\}_{k=1}^{K}$ denote the target-object queries, where

$$q_k = (u_k, j_k, \mathbf{b}_k, \ell_k)$$

specifies an object identifier u_k, a designated view I_{j_k}, its 2D bounding box $\mathbf{b}_k$ in that view, and its category label ℓ_k. Given $\mathcal{V}$ and $\mathcal{Q}$, the model predicts

$$\widehat{\mathcal{S}} = F(\mathcal{V}, \mathcal{Q}) = \{(u_k, \hat{\mathbf{c}}_k, \hat{\mathbf{s}}_k, \hat{\psi}_k)\}_{k=1}^{K}, \tag{1}$$

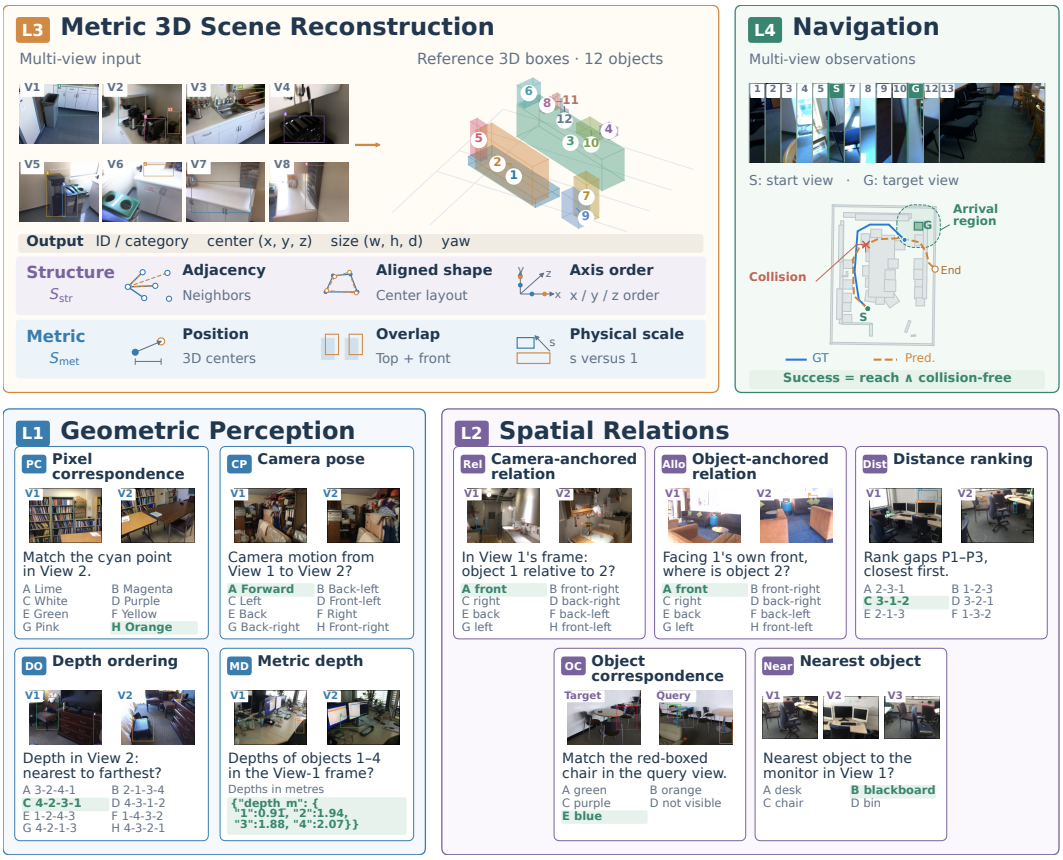


Figure 2: Examples across the four levels of RECONSPATIAL, with reconstruction scoring schematics and navigation annotations on the shown scene. Bold green options mark reference answers. Prompts are abbreviated; complete examples appear in Appendix B.5.

where $\hat{\mathbf{c}}_k \in \mathbb{R}^3$ and $\hat{\mathbf{s}}_k \in \mathbb{R}^3_+$ denote the center and side lengths of the predicted 3D box, and $\hat{\psi}_k$ denotes its yaw about the gravity axis. All quantities are expressed in metres in a gravity-aligned frame whose horizontal orientation is anchored to I_1. Its origin is the vertical projection of the reference-camera center onto the support surface.

Providing object identifiers, category labels, and image regions controls for open-set detection while preserving the need to associate targets across views and predict a mutually compatible scene. Oriented 3D boxes make the inferred geometry directly measurable without assuming that the model internally represents the scene in the same form. Predictions are matched to queried objects by identifier under the fixed parsing and missing-output protocol detailed in Appendix C.

3.2 COMPLEMENTARY DIMENSIONS OF SPATIAL FIDELITY

A reconstruction may preserve relative scene organization while remaining inaccurate in physical scale and camera-anchored placement. We therefore evaluate spatial fidelity along two complementary dimensions: *Structure*, which measures relative organization, and *Metric*, which measures camera-anchored placement and physical scale:

$$S_{\text{ord}} = \tfrac{1}{3}(S_{\text{lr}} + S_{\text{fb}} + S_{\text{up}}), \tag{2}$$

$$S_{\text{str}} = \tfrac{1}{3}(S_{\text{adj}} + S_{\text{shape}} + S_{\text{ord}}), \tag{3}$$

$$S_{\text{met}} = \tfrac{1}{3}(S_{\text{pos}} + S_{\text{viewIoU}} + S_{\text{scale}}). \tag{4}$$

All component scores lie in $[0, 1]$. Structure terms are additionally calibrated against a random-layout baseline, mapping chance agreement to zero.

Structure combines alignment-invariant layout agreement with camera-anchored directional consistency. Neighbour-graph agreement (S_{adj}) evaluates horizontal adjacency, while S_{shape} measures the relative configuration of object centers after planar similarity alignment; both terms are invariant to planar translation, rotation, and uniform scale. In contrast, S_{ord} measures left–right, front–back, and vertical ordering in the anchored frame, retaining directional constraints that similarity alignment removes.

Metric evaluates placement and absolute scale in the anchored frame. S_{pos} measures 3D center-localization accuracy across multiple distance thresholds, while S_{viewIoU} averages matched-box overlap in orthographic ground-plane (x, z) and front-plane (x, y) projections, jointly reflecting position, extent, and orientation. Both components use a single scale fit; no rotation or translation is fitted, preserving sensitivity to errors in the anchored coordinate frame. The separate S_{scale} term penalizes deviation of the fitted scale from one, preventing scale alignment from concealing global metric miscalibration. The three Metric terms contribute equally. Figure 2 illustrates the six scoring components; Appendix C provides full formulas and calibration details.

3.3 Diagnostic Hierarchy and Evaluation Domains

The reconstruction suite comprises 840 instances across indoor, tabletop, and outdoor environments (Figure 1(b)). Scores are first averaged within each domain and then across the three domain means, preventing domains with larger sample counts from dominating the aggregate score.

The benchmark tasks are designed so that relevant evidence is distributed across multiple views. Correspondence and camera-pose tasks require comparisons across observations; depth and relation queries place objects from different views into a shared reference frame; and reconstruction and navigation require combining partial observations into a coherent scene representation. Questions and reference answers are constructed from dataset-provided geometry, instance annotations, and geometrically derived labels (Appendix B).

Because an aggregate reconstruction score cannot isolate constituent failure modes, we pair holistic reconstruction with nine atomic diagnostics organized into two complementary task families (Figure 2). *Geometric Perception* examines cross-view correspondence, camera pose, depth ordering, and metric depth. *Spatial Relations* assesses reasoning across reference frames through camera- and object-anchored directions, distance ranking, object correspondence, and proximity. Detailed task formulations appear in Appendix A.

To evaluate spatial planning as an independent downstream task, we include 346 offline navigation episodes in single-room environments. Given multi-view observations, a designated starting viewpoint, and a visual target, the model directly predicts an egocentric trajectory. Success requires reaching the target without colliding with annotated obstacles, accepting any valid path rather than enforcing agreement with a single reference trajectory (Appendix F).

Together, these tasks benchmark constituent spatial judgments, joint 3D scene reconstruction, and goal-directed planning.

4 Empirical Analysis

In this section, we evaluate 18 frontier LVLMs across the four levels of ReconSpatial. We examine the overall benchmark landscape, the disparity between Structure and Metric fidelity, and the gap between atomic capabilities and joint scene reconstruction.

4.1 Evaluation setup and model cohort

The cohort comprises GPT-6-Astra (OpenAI, 2026), GPT-5.6-Sol, and GPT-5.5; Gemini-3.1-Pro (Comanici et al., 2025), Gemini-3.5-Flash (Google, 2026), Gemini-3.7-Flash, and Gemini-3.8-Flash; Claude-Opus-4.7, Claude-Opus-4.8, Claude-Opus-5, and Claude-Fable-5; Qwen3.7-Plus (Bai et al., 2025) and Qwen3.8-Max (Qwen Team, 2026); Kimi-K3 (Team et al., 2026); Doubao-Seed-2.1-Pro (ByteDance Seed, 2026); MiniMax-M3 (MiniMax, 2026); Grok-4.6; and GLM-5.3-Flash (Z.ai, 2026). All models are evaluated under identical prompts and execution settings (Appendix C). The Overall score equally weights the four capability levels—Geometric Perception, Spatial Relations,

Table 1: Main results for eighteen frontier LVLMs. Atomic-task abbreviations are defined in Figure 2; Reconstruction reports three-domain Structure (S), Metric (M), and their mean. Scores are percentages; Navigation uses scale-normalized paths and Overall equally averages the four levels. Best and second-best values are bold and underlined, respectively.

Model	Geometric Perception					Spatial Relations						Reconstruction			Navigation	Overall
	PC	CP	DO	MD	Avg.	Rel	Allo	Dist	OC	Near	Avg.	S	M	Avg.	Avg.	Avg.
GPT-6-Astra	**87.9**	**70.8**	**89.1**	**53.6**	**75.3**	**53.3**	50.8	61.4	72.7	**74.0**	**62.4**	**81.4**	**58.7**	**70.0**	**30.3**	**59.5**
Qwen3.8-Max	74.0	56.2	83.3	47.1	65.2	50.2	**60.3**	54.3	67.7	61.5	58.8	68.5	36.7	52.6	17.1	48.4
Kimi-K3	63.3	50.6	72.7	48.0	58.6	36.1	42.9	49.2	65.2	52.1	49.1	68.9	32.7	50.8	21.7	45.1
Gemini-3.8-Flash	64.2	54.1	70.9	49.5	59.7	41.9	23.8	**65.0**	66.7	65.6	52.6	66.4	37.6	52.0	12.1	44.1
GPT-5.6-Sol	56.7	50.9	72.7	41.8	55.5	42.5	34.9	34.0	53.0	46.9	42.3	71.3	41.5	56.4	21.7	44.0
Claude-Opus-5	58.1	51.9	74.6	29.6	53.6	48.0	47.6	54.3	47.0	59.4	51.3	68.1	38.9	53.5	17.3	43.9
Claude-Fable-5	55.4	46.4	69.1	38.8	52.4	39.2	44.4	45.7	42.4	65.6	47.5	68.7	39.4	54.1	16.2	42.5
Gemini-3.7-Flash	63.7	52.7	58.2	46.6	55.3	42.3	25.4	53.3	69.7	55.2	49.2	53.7	33.6	43.6	16.2	41.1
GLM-5.3-Flash	44.6	45.9	74.6	49.2	53.6	31.7	23.8	50.8	47.0	50.0	40.7	68.3	37.2	52.8	15.6	40.7
GPT-5.5	40.9	45.8	74.6	46.1	51.9	34.4	41.3	40.6	51.5	45.8	42.7	65.0	28.6	46.8	17.9	39.8
Doubao-Seed-2.1-Pro	47.9	46.0	63.6	45.3	50.7	33.0	7.9	44.2	**74.2**	56.2	43.1	63.8	30.2	47.0	17.1	39.5
Grok-4.6	32.1	45.9	80.0	43.6	50.4	33.9	41.3	42.1	59.1	50.0	45.3	63.3	29.4	46.4	15.6	39.4
Gemini-3.1-Pro	44.6	43.6	60.0	36.3	46.1	28.2	38.1	26.9	43.9	36.5	34.7	62.3	29.6	45.9	11.8	34.6
Gemini-3.5-Flash	48.4	46.9	43.6	36.9	44.0	33.9	31.8	37.1	63.6	49.0	43.1	49.1	32.9	41.0	10.1	34.5
Qwen3.7-Plus	37.7	41.3	65.5	33.4	44.5	30.4	46.0	32.5	53.0	33.3	39.1	52.7	19.6	36.1	13.3	33.2
Claude-Opus-4.8	29.8	38.0	56.4	31.3	38.9	26.1	25.4	37.6	28.8	35.4	30.7	55.5	24.3	39.9	9.8	29.8
Claude-Opus-4.7	29.3	36.6	40.0	34.1	35.0	29.2	23.8	31.0	40.9	26.0	30.2	53.9	21.1	37.5	10.1	28.2
MiniMax-M3	30.7	37.5	50.9	40.5	39.9	25.6	11.1	26.4	18.2	28.1	21.9	50.7	18.2	34.4	9.2	26.4

Reconstruction, and Navigation—and Reconstruction equally averages Structure and Metric across the three physical domains, preventing split sizes from biasing aggregate results.

4.2 Benchmark Landscape and the Structure–Metric Disparity

Table 1 illustrates the overall benchmark landscape. GPT-6-Astra sets the performance frontier, leading all four levels with 59.5 Overall, followed by Qwen3.8-Max at 48.4. Across capability tiers, runner-up rankings vary: Qwen3.8-Max ranks second in both atomic levels, while GPT-5.6-Sol places second in reconstruction and ties with Kimi-K3 for second in navigation. Farther down the leaderboard, Claude-Fable-5 ranks third in reconstruction (54.1) but drops to seventh Overall (42.5). Close aggregate scores thus frequently mask divergent competencies across local geometric reasoning, scene reconstruction, and planning. Across all models, navigation remains universally difficult: even GPT-6-Astra succeeds in only 30.3% of episodes under scale-normalized evaluation, motivating our map-substitution experiments to probe the origins of planning failures in Section 5.2.

The most consistent empirical trend is the disparity between relative layout recovery and camera-anchored metric placement. Structure exceeds Metric in all 54 evaluated model–domain combinations (Figure 3); for instance, GPT-6-Astra achieves 81.4% on Structure but only 58.7% on Metric. For every model, the scene-cluster bootstrap interval for this cross-domain gap remains strictly above zero (Appendix D). This pattern indicates that models recover relative layout far more reliably than physical scale and camera-anchored placement. The gap also persists under our tested scorer controls and alternative component weights (Appendices C and E.3).

4.3 Relating Atomic Capabilities to Holistic Reconstruction

To isolate task-specific associations from general model capability, we compute partial correlations between individual atomic skills and reconstruction metrics (Appendix D). The resulting associations align closely with the geometric requirements of multi-view synthesis: camera-pose estimation correlates strongly with projected box overlap (partial $r = 0.62$), reflecting the critical role of inter-view registration in establishing a consistent 3D consensus. Similarly, depth-ordering accuracy tracks front–back layout consistency ($r = 0.63$), cross-view pixel correspondence correlates with aggregate Metric scores ($r = 0.46$), and camera-anchored direction judgments track 3D position

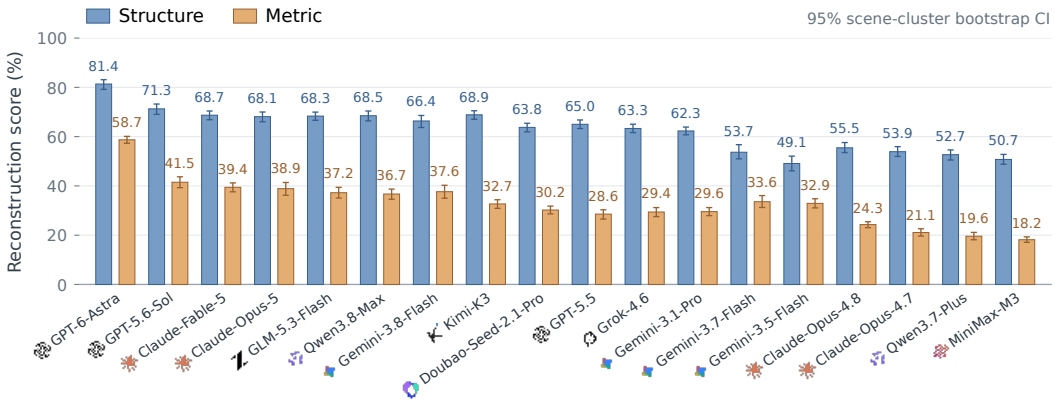


Figure 3: Structure and Metric reconstruction scores across eighteen models, ordered by descending mean score. Paired bars show cross-domain means; error bars denote 95% scene-cluster bootstrap intervals. Structure consistently exceeds Metric across the model cohort.

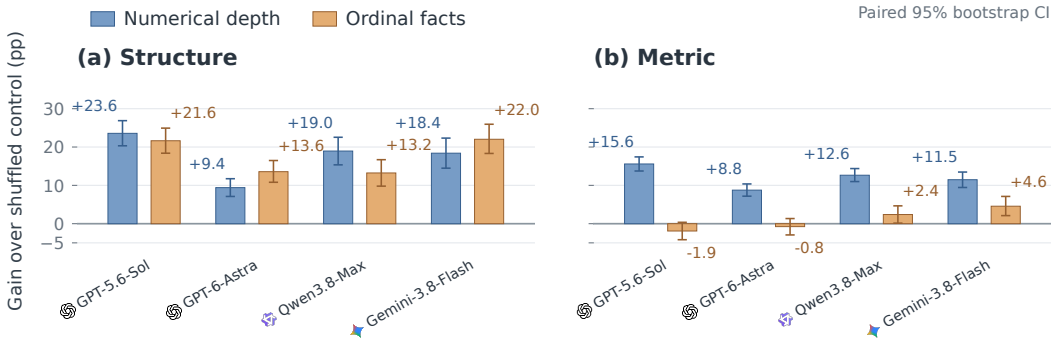


Figure 4: Geometric interventions on reconstruction. Gains are relative to shuffled controls; error bars show 95% bootstrap CIs.

accuracy ($r = 0.34$). Because observational correlations alone cannot establish causality, Section 5 further examines whether supplying explicit geometric information improves reconstruction.

5 Diagnosing and Improving Spatial Understanding

The Structure–Metric disparity raises two core questions: what information bottlenecks metric scene reconstruction, and whether recovered 3D geometry improves downstream planning. Below, we investigate these questions through targeted information interventions, downstream map substitution, qualitative error analysis, and cross-task spatial fine-tuning.

5.1 Information Bottlenecks in Metric Reconstruction

Across four models, numerical-depth information improves Metric more than ordinal relations, while both cues similarly enhance Structure (Figure 4). This contrast demonstrates that relational consistency alone does not provide the precision required for accurate metric reconstruction. Further probes on GPT-5.6-Sol reveal no reliable depth improvement from increasing the visual context from two to ten views or providing one object's true physical size (Appendix G). Under these tested conditions, richer visual evidence does not readily translate into more accurate metric estimates, highlighting a limitation in converting multi-view observations into calibrated geometry.

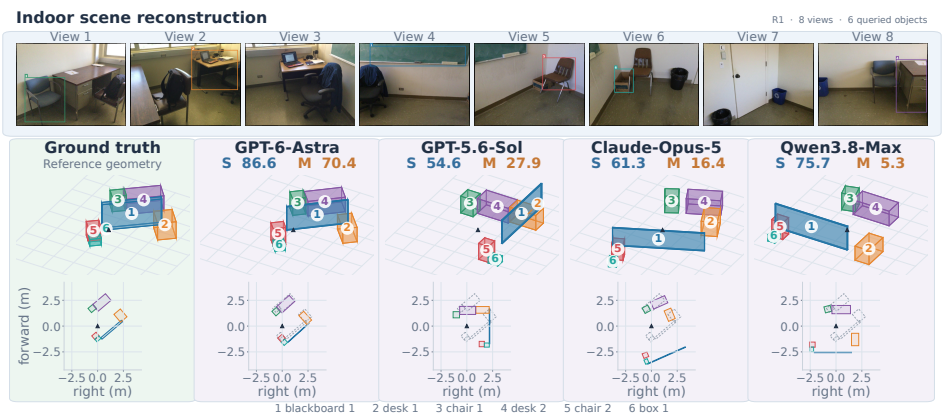


Figure 5: Indoor scene reconstruction (R1). Input views, predicted 3D boxes, and ground-plane footprints share object colors. Dashed footprints denote GT; triangles mark the reference origin. S/M denote Structure/Metric (0–100). All models share metric axes without alignment or rescaling.

5.2 3D Maps as Planning Intermediates

We further leverage the navigation task to evaluate 3D reconstruction as an intermediate representation for planning. *Images* denotes direct image-only planning. In *Self-map*, the model first reconstructs a map and then predicts a path; *GT map* supplies ground-truth boxes instead. *GT + target ID* additionally identifies the target entry in the ground-truth map (Table 2). Across evaluated models, ground-truth maps substantially improve navigation success, whereas self-predicted maps yield smaller, model-dependent gains. These gains show that errors in scene construction account for a substantial part of navigation failures. Furthermore, explicit target-ID annotation provides secondary, model-dependent benefits (Appendix F).

Table 2: Navigation success (%) under 3D box map substitution, without path rescaling.

Model	Images	Self-map	GT map	GT + target ID
GPT-5.6-Sol	15.6	15.6	85.3	**90.5**
GPT-6-Astra	22.5	35.5	97.1	**97.4**
Qwen3.8-Max	9.5	11.0	66.2	**78.6**
Gemini-3.8-Flash	11.8	18.2	**82.1**	79.8

5.3 Qualitative Analysis

Inspecting predictions on identical inputs exposes geometric discrepancies that aggregate metrics obscure. Figure 5 visualizes multi-view reconstructions for an indoor scene within a shared metric coordinate frame. While all evaluated models successfully identify the targets and recover plausible relative arrangements, their metric geometries diverge substantially. Notably, Qwen3.8-Max preserves recognizable object clusters (Structure: 75.7), yet collapses in camera-anchored placement (Metric: 5.3) due to a 1.82 m global drift. This case makes the physical separation between relational coherence and metric accuracy directly inspectable. Additional tabletop and navigation failure analyses are provided in Appendix I.

5.4 Learning from Cross-View Spatial Supervision

Beyond diagnosis, we examine whether cross-view spatial supervision transfers to broader multimodal reasoning. We fine-tune Qwen3.6-Flash on up to 50k spatial examples spanning Geometric Perception, Spatial Relations, and Reconstruction, without general-domain training data (Appendix H). On external evaluations, the 50k Spatial SFT model improves nine of ten spatial benchmarks and all eight general benchmarks (Table 3), led by VisualPuzzles (+7.9) and BabyVision (+3.6). These results suggest that cross-view geometric consistency provides a transferable inductive bias for multimodal reasoning. Across data scales, external transfer saturates early: spatial and general averages already reach 64.08 and 74.30 with 10k examples, exhibiting minimal further change through 20k and 50k

Table 3: Spatial supervised fine-tuning (SFT) of Qwen3.6-Flash on **(a)** spatial, **(b)** general, and **(c)** RECONSPATIAL evaluations. The external and RECONSPATIAL panels report all three spatial-data doses and their starting-checkpoint reference. Scores are percentages and column-best values are bold; GP/SR and S/M denote Geometric Perception/Spatial Relations and Structure/Metric.

(a) Spatial benchmarks

Variant	SITE	BLINK	View Spatial	ERQA	Omni Spatial	3DSR	MMSI	Emb Spatial	SITE Video	Ref Spatial	Avg.
Base	56.48	68.15	48.72	58.25	57.99	64.41	38.50	82.72	71.26	63.36	60.98
SFT 10k	63.12	73.82	52.92	59.75	**62.95**	67.09	39.82	83.74	72.12	**65.46**	64.08
SFT 20k	64.71	**76.62**	53.57	60.75	62.56	**68.66**	42.42	84.18	**73.22**	63.45	65.01
SFT 50k	**65.54**	75.42	**54.20**	**62.75**	62.30	68.53	**42.50**	**84.56**	72.84	62.64	**65.13**

(b) General benchmarks

	Multimodal knowledge		Visual reasoning		Charts and diagrams		Scene / perceptual reliability		
Variant	MMMU	MMBench CN	Visual Puzzles	Baby Vision	CharXiv	AI2D	Real WorldQA	Hallusion Bench	Avg.
Base	80.89	89.94	53.42	24.23	74.70	92.26	82.75	78.55	72.09
SFT 10k	81.56	90.79	**63.36**	**30.67**	76.30	92.55	82.88	76.34	74.30
SFT 20k	82.00	90.94	62.24	27.06	**78.10**	**92.78**	83.01	78.13	74.28
SFT 50k	**82.56**	**91.56**	61.30	27.84	77.00	**92.78**	**83.66**	**79.92**	**74.58**

(c) RECONSPATIAL

Configuration	GP	SR	Reconstruction			Nav.	Overall
	Avg.	Avg.	S	M	Avg.		
Base (reference)	38.44	20.03	42.52	19.48	31.00	8.10	24.39
Spatial SFT (10k)	49.29	35.91	50.11	26.48	38.29	**11.80**	33.82
Spatial SFT (20k)	50.43	40.87	52.70	26.33	39.52	8.40	34.80
Spatial SFT (50k)	**54.22**	**43.37**	**54.74**	**26.60**	**40.67**	10.40	**37.17**

On RECONSPATIAL, in contrast, performance compounds steadily with data scale (rising from 24.39 to 37.17 Overall), driven by monotonic improvements across geometric perception, spatial relations, and 3D reconstruction. Downstream navigation, however, remains below 12% with non-monotonic variations. These results show that cross-view spatial supervision provides a transferable benefit for both local perception and holistic reconstruction, while reliably translating these perceptual gains into downstream planning remains an open challenge.

6 CONCLUSION AND LIMITATIONS

In this work, RECONSPATIAL establishes multi-view 3D scene reconstruction as a principled benchmark connecting local spatial perception with holistic scene coherence and downstream planning. Across 18 frontier LVLMs, our evaluations uncover a persistent gap between relative scene structure and metric reconstruction. Diagnostic analyses show that relational consistency alone cannot guarantee metric accuracy, while downstream map substitution demonstrates that navigation failures stem primarily from perceptual geometry rather than planning logic. Furthermore, cross-view spatial fine-tuning confers broad transfer to both spatial and general vision-language benchmarks.

Limitations and Future Work. Several limitations outline avenues for future work. First, physical objects are represented via oriented 3D bounding boxes, omitting fine-grained surface meshes, articulated parts, and continuous free-space occupancy. Second, our navigation evaluation focuses on offline, open-loop planning within single rooms to decouple execution noise, leaving closed-loop recovery and multi-room exploration unexamined. Finally, spatial fine-tuning was evaluated on a single model family. Scaling these evaluations to broader architectures and interactive physical interaction represents a critical frontier for building dependable embodied spatial intelligence.

ETHICS STATEMENT

Data Sourcing and Privacy. RECONSPATIAL builds upon existing open research datasets (Appendix B). All asset curation and redistribution strictly comply with the original licenses and permissible use policies. To protect privacy, all real-world scenes have been screened to ensure they contain no personally identifiable information.

Broader Impact and Safety. This benchmark aims to advance foundational 3D spatial perception and reasoning. All navigation tasks are evaluated strictly offline using pre-recorded observations and maps, involving no physical robotic actuation or safety risks. While enhanced metric accuracy is essential for building more dependable embodied agents, real-world deployment requires independent, rigorous safety verification and failsafe mechanisms.

AI USE STATEMENT

We used LLMs to assist with manuscript editing, code development and figure preparation. Separately, Qwen3.8-Max was employed to generate reasoning traces and target responses for spatial fine-tuning (Appendix H). All AI-assisted outputs, code, and synthesized data were manually verified by the authors, who take full responsibility for all methods, results, and claims presented in this work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, et al. AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems. arXiv preprint arXiv:2503.06669, 2025. URL https://arxiv.org/abs/2503.06669.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

ByteDance Seed. Seed 2.1 Pro, 2026. URL https://seed.bytedance.com/en/seed2_1.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://arxiv.org/abs/1903.11027.

Liang Chen, Weichu Xie, Yiyan Liang, et al. BabyVision: Visual Reasoning Beyond Language. arXiv preprint arXiv:2601.06521, 2026. URL https://arxiv.org/abs/2601.06521.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017. URL https://arxiv.org/abs/1702.04405.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking Spatial Understanding for Embodied Tasks with Large Vision-Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL https://arxiv.org/abs/2406.05756.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding, 2026. URL https://arxiv.org/abs/2604.05015.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal Large Language Models Can See but Not Perceive. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2404.12390.

Gemini Robotics Team. Gemini Robotics: Bringing AI into the Physical World. arXiv preprint arXiv:2503.20020, 2025. URL https://arxiv.org/abs/2503.20020. Source of the ERQA (Embodied Reasoning Question Answering) benchmark.

Google. Gemini 3.5 Flash, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2310.14566.

Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 880–888, 2017.

Yining Hong, Jiageng Liu, Han Yin, Manling Li, Leonidas Guibas, Fei-Fei Li, Jiajun Wu, and Yejin Choi. ESI-Bench: Towards embodied spatial intelligence that closes the perception-action loop. *arXiv preprint arXiv:2605.18746*, 2026.

Yibin Huang, Wang Xu, Wanyue Zhang, Helu Zhi, Jingjing Huang, Yangbin Xu, Yangang Sun, Conghui Zhu, and Tiejun Zhao. Video2Layout: Recall and reconstruct metric-grounded cognitive map for spatial reasoning. *arXiv preprint arXiv:2511.16160*, 2025.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, 2026.

Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea Open-World Dataset and G0 Dual-System VLA Model. arXiv preprint arXiv:2509.00576, 2025. URL https://arxiv.org/abs/2509.00576.

Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram Is Worth A Dozen Images. In *European Conference on Computer Vision (ECCV)*, pp. 235–251, 2016. URL https://arxiv.org/abs/1603.07396.

Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify Anything: Scaling Indoor 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22225–22233, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lazarow_Cubify_Anything_Scaling_Indoor_3D_Object_Detection_CVPR_2025_paper.html.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025a.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang, Yuwei Wu, Tong He, Jiangmiao Pang, Yu Qiao, Yunde Jia, and Kaipeng Zhang. Sekai: A Video Dataset towards World Exploration. arXiv preprint arXiv:2506.15675, 2025b. URL https://arxiv.org/abs/2506.15675.

Parker Liu, Chenxin Li, Zhengxin Li, Yipeng Wu, Wuyang Li, Zhiqin Yang, Zhenyuan Zhang, Yunlong Lin, Sirui Han, and Brandon Y. Feng. IR3D-Bench: Evaluating vision-language model scene understanding as agentic inverse rendering. *arXiv preprint arXiv:2506.23329*, 2025a.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 29982–30009, 2025b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player? In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2307.06281.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso De Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934. IEEE, 2025.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. OpenEQA: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498. IEEE, 2024.

MiniMax. MiniMax-M3, 2026. URL https://www.minimax.io/models/text/m3.

OpenAI. GPT-6 Astra: A new generation of intelligence, September 2026. URL https://openai.com/index/gpt-6-astra/.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Lukas Petersson, Axel Backlund, Axel Wennström, Hanna Petersson, Callum Sharrock, and Arash Dabiri. Blueprint-Bench: Comparing spatial intelligence of LLMs, agents and image models, 2025. URL https://arxiv.org/abs/2509.25229.

Qwen Team. Qwen3.8-Max: A New Bar for Coding and Cowork, 2026. URL https://qwen.ai/blog?id=qwen3.8.

Thomas Sch"ops, Johannes L. Sch"onberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3260–3269, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Schops_A_Multi-View_Stereo_CVPR_2017_paper.html.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. VisualPuzzles: Decoupling Multimodal Reasoning Evaluation from Domain Knowledge. arXiv preprint arXiv:2504.10342, 2025. URL https://arxiv.org/abs/2504.10342.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A Dataset for Robot Learning at Scale. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*, 2023. URL https://arxiv.org/abs/2308.12952.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2642–2651, 2019. URL https://arxiv.org/abs/1901.02970.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. EmbodiedScan: A Holistic Multi-Modal 3D Perception Suite Towards Embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a. URL https://arxiv.org/abs/2312.16170.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. SITE: Towards spatial intelligence thorough evaluation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9058–9069. IEEE, 2025.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 113569–113697, 2024b. URL https://arxiv.org/abs/2406.18521.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Advances in Neural Information Processing Systems*, 2025a.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, Shichao Fan, Xinhua Wang, Fei Liao, Zhen Zhao, Guangyu Li, Zhao Jin, Lecheng Wang, Jilei Mao, Ning Liu, Pei Ren, Qiang Zhang, Yaoxu Lyu, Mengzhen Liu, Jingyang He, Yulin Luo, Zeyu Gao, Chenxuan Li, Chenyang Gu, Yankai Fu, Di Wu, Xingyu Wang, Sixiang Chen, Zhenyu Wang, Pengju An, Siyuan Qian, Shanghang Zhang, and Jian Tang. RoboMIND: Benchmark on Multi-embodiment Intelligence Normative Data for Robot Manipulation. In *Robotics: Science and Systems (RSS)*, 2025b. doi: 10.15607/RSS.2025.XXI.152. URL https://arxiv.org/abs/2412.13877.

xAI. RealWorldQA. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Real-world spatial-understanding QA benchmark released by xAI with Grok-1.5V; CC BY-ND 4.0.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *Robotics: Science and Systems (RSS)*, 2018. URL https://arxiv.org/abs/1711.00199.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, 2026a.

Yixuan Yang, Zhen Luo, Wanshui Gan, Jinkun Hao, Junru Lu, Jinghao Yan, Zhaoyang Lyu, and Xudong Xu. Code-as-Room: Generating 3D Rooms from Top-Down View Images via Agentic Code Synthesis. arXiv preprint arXiv:2605.18451, 2026b. URL https://arxiv.org/abs/2605.18451.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2308.11417.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.16502.

Z.ai. GLM-5.3-Flash, 2026. URL https://z.ai/blog/glm-5.3-flash.

Enhan Zhao, Wei Wu, Yuanrui Zhang, Xueliang Zhao, and Di He. Ouroboros-Spatial: Closing the data-model loop for spatial reasoning, 2026a. URL https://arxiv.org/abs/2606.11719.

Yanpeng Zhao, Wentao Ding, Hongtao Li, Baoxiong Jia, and Zilong Zheng. ESPIRE: A diagnostic benchmark for embodied spatial reasoning of vision-language models, 2026b. URL https://arxiv.org/abs/2603.13033.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2506.04308.

Qihui Zhu, Shouwei Ruan, Xiao Yang, Hao Jiang, Yao Huang, Shiji Zhao, Hanwei Fan, Hang Su, and Xingxing Wei. Mind over space: Can multimodal large language models mentally navigate? *arXiv preprint arXiv:2603.21577*, 2026.

A Detailed Task Formulations

The atomic tasks assess spatial skills involved in multi-view 3D scene reconstruction. This appendix specifies their inputs, reference frames, and answer spaces; all scores are computed analytically from annotated multi-view data. Table 4 reports domains, sources, counts, and views.

MCQ reference accuracy averages the inverse number of options per item: 16.8% for pixel correspondence, 15.6% for camera pose, 12.5% for depth ordering and both directional-relation tasks, 16.7% for distance ranking, 20.7% for object correspondence, and 25.0% for nearest object. Metric depth uses mean accuracy over relative-error thresholds, with a constant-depth reference of 31.2%.

A.1 Geometric Perception

Pixel correspondence. The input contains two views and a query marker in the first. The second view contains five to eight colored candidate markers. The model selects the marker corresponding to the same physical surface point. Items span indoor, outdoor, and tabletop scenes and are scored by exact multiple-choice accuracy. Candidate locations are sampled around the projected correspondence so that image proximity alone does not identify the answer.

Camera pose. Two views are presented without intrinsics or field of view. Translation is quantized into eight egocentric directions relative to the first camera; rotation is represented by its sense and magnitude. The reported unit score is the mean of translation and rotation accuracy. This discrete formulation avoids conflating geometric understanding with unstable free-form numeric formatting.

Table 4: Data composition of RECONSPATIAL across L1–L4. Counts are task instances, not distinct physical scenes; views denote input frames. ScanNet and CA-1M are pooled within the indoor domain.

Task	Domain	N	Views
L1: Geometric Perception			
Pixel correspondence	Indoor	56	2
	Tabletop	118	2
	Outdoor	42	2
Camera pose	Indoor, outdoor	361	2
Depth ordering	Indoor	55	2
Metric depth	Indoor	107	2
L2: Spatial Relations			
Camera-anchored relation	Indoor	227	2
Object-anchored relation	Indoor	63	2
Distance ranking	Indoor	197	2
Object correspondence	Indoor	66	2–6
Nearest object	Indoor	96	3
L3: Scene Reconstruction			
Oriented-box reconstruction	Indoor	475	2 / 8
	Tabletop	89	2
	Outdoor	276	6
L4: Navigation			
Offline path prediction	Indoor	346	11–23

Data sources

Task	Domain	Data sources
Pixel correspondence	Indoor	ScanNet (Dai et al., 2017).
	Tabletop	AgiBot World (AgiBot-World-Contributors et al., 2025), BridgeData V2 (Walke et al., 2023), in-house data, RDT-1B-associated recordings (Liu et al., 2025b), RoboMIND (Wu et al., 2025b), and Galaxea Open-World (Jiang et al., 2025).
	Outdoor	Sekai (Li et al., 2025b).
Camera pose	Indoor, outdoor	ScanNet, ScanNet++ (Yeshwanth et al., 2023), ETH3D (Sch"ops et al., 2017), and Sekai.
Other tasks	Indoor	ScanNet and CA-1M (Lazarow et al., 2025). Object-anchored relation and navigation use ScanNet only.
Scene reconstruction	Tabletop	YCB-V (Xiang et al., 2018), T-LESS (Hodaň et al., 2017), HomebrewedDB (Kaskman et al., 2019), and NOCS (Wang et al., 2019).
	Outdoor	nuScenes (Caesar et al., 2020).

Depth ordering. Four objects are designated by boxes across two views. The model orders their centres by depth along the second camera's viewing axis, expressing all queried objects in the same reference frame. The output is one of eight candidate permutations and is scored by exact accuracy.

Metric depth. The model receives two views with four named, boxed objects and predicts each object's depth in metres along a designated reference camera axis. At least one object is marked in the other view, requiring its depth to be expressed in the reference camera frame. The output is a JSON mapping from object identifiers to real-valued depths. We report the mean accuracy at relative-error thresholds $\{5, 10, 20, 40\}\%$, rather than an options-based accuracy.

A.2 Spatial Relations

Camera-anchored relation. Two objects appear in different views. The model adopts the first camera's orientation and classifies the direction of one object relative to the other into eight horizontal sectors. The reference-frame requirement distinguishes cross-view perspective taking from a comparison of image-plane positions.

Object-anchored relation. The reference frame is defined by an anchor object's own front direction rather than by a camera. Anchors are restricted to categories with a meaningful front, and candidate examples are manually checked for orientation ambiguity. The answer uses the same eight-direction output space as the camera-anchored task.

Distance ranking. Each item specifies three object pairs whose members may occur in different views. The model ranks the pairs by their Euclidean 3D separation. The six possible permutations form the answer space. This task measures inter-object metric structure and differs from depth ordering, which is defined along a camera axis.

Object correspondence (OC). This task uses same-category, size-similar objects that cannot be reliably distinguished by their names. In the correspondence variant, the first image marks a target with a red box and the query view marks candidates with distinct colors; the model selects the same instance or reports that it is not visible. In the elimination variant, context views assign colors to several instances and the model identifies which one is absent from the query view. The two variants are micro-averaged into one score.

Nearest object. A reference object and four candidates are distributed across three views. The model selects the candidate with the smallest 3D distance to the reference. Construction records the true distance margin and screens simple image-space cues such as box area and image center. The task is scored by four-way accuracy.

B Task Construction and Examples

ReconSpatial is constructed around cross-view scenes: answering a query requires combining observations from multiple viewpoints, since no single view provides the complete task evidence. This principle takes different forms across the suite, from establishing correspondences between views to expressing objects in a common reference frame and recovering a shared layout for reconstruction and navigation. We organize each source into scene records linking images, camera geometry, and available point- or object-level annotations, from which we derive task-specific questions and reference answers. The construction procedures are described below, followed by a compact collection of examples with visual inputs, complete prompts, and reference answers.

B.1 Geometric Perception

Pixel correspondence. The data cover indoor scenes from ScanNet Dai et al. (2017), outdoor scenes from Sekai Li et al. (2025b), and robot observations from AgiBot World AgiBot-World-Contributors et al. (2025), BridgeData V2 Walke et al. (2023), RoboMIND Wu et al. (2025b), Galaxea Open-World Jiang et al. (2025), RDT-1B-associated recordings Liu et al. (2025b), and in-house data. Robot examples pair distinct viewpoints from head-mounted, wrist-mounted, and external third-person cameras, according to the cameras available in each source. Existing correspondence annotations are used where available. For robot observations, we detect target regions and use MASt3R Leroy et al. (2024) to obtain cross-view matches and 3D point maps. Candidate matches are retained based on the spatial agreement of their associated 3D points. A source point, its correspondence, and distractor locations are then rendered as query and candidate markers, yielding a multiple-choice question about the same physical point across views.

Camera pose. We use images and associated camera poses from ScanNet Dai et al. (2017), ScanNet++ Yeshwanth et al. (2023), ETH3D Sch"ops et al. (2017), and Sekai Li et al. (2025b). For each selected view pair, the poses define the relative camera transformation, from which we derive

translation-direction and rotation labels and construct answer choices. Both variants require relating the two camera frames; translation is queried as a direction, accommodating sources with poses defined only up to scale.

Depth ordering and metric depth. Both tasks use ScanNet scenes with EmbodiedScan annotations Dai et al. (2017); Wang et al. (2024a) and CA-1M Lazarow et al. (2025). We select target objects across views and transform their annotated 3D centres into a designated reference camera frame. Their coordinates along its viewing axis define the reference depths, while image regions identify the queried objects. Depth ordering converts these depths into a near-to-far ordering and alternative permutations. Every adjacent pair in the sorted sequence must differ by at least $0.3\,\mathrm{m}$ to avoid ambiguous orderings. Metric depth instead uses the reference-axis depths as numerical answers, without imposing this ordering-specific separation criterion. Both tasks require relating the target observations to the designated reference view.

B.2 Spatial Relations

Spatial relation tasks draw on object-level geometry from ScanNet Dai et al. (2017); Wang et al. (2024a) and CA-1M Lazarow et al. (2025); object-anchored directions use ScanNet objects with a well-defined orientation. Scene records associate object centres, orientations, instance identities, and visibility across views. Queries are formed by distributing the required object evidence across views and deriving directional, distance, or identity relations from these records. This annotation basis also supports joint 3D box reconstruction: relation tasks isolate local judgments and reference-frame reasoning, whereas reconstruction requires recovering the object layout jointly.

Camera-anchored relation. Object pairs are selected from different views so that neither view contains both targets. Their relative displacement is expressed in the designated camera frame and mapped to a directional label, which determines the correct answer among the candidate relations.

Object-anchored relation. An anchor object with a meaningful front direction defines a local reference frame. The position of a target observed in another view is transformed into this frame to determine its direction relative to the anchor, requiring the target location and anchor orientation to be combined across views.

Distance ranking. We form object pairs whose endpoints are observed across different views and compute their Euclidean 3D centre distances. Sorting these distances gives the reference ranking; alternative permutations provide the answer choices.

Object correspondence. We select instances of similar category and size and use their identities and per-view visibility to associate them across observations. Matching questions ask which candidate corresponds to a target in another view; missing-instance questions ask which context instance is absent from the query view. Coloured markers identify the candidates, and instance annotations determine the answer in both variants.

Nearest object. An anchor and candidate objects are distributed across views. Their annotated 3D centres determine the anchor-to-candidate distances, and the candidate with the smallest distance supplies the reference answer.

B.3 Scene Reconstruction

Joint 3D box reconstruction. Indoor scenes come from ScanNet Dai et al. (2017); Wang et al. (2024a) and CA-1M Lazarow et al. (2025); tabletop scenes from YCB-Video Xiang et al. (2018), T-LESS Hodaň et al. (2017), HomebrewedDB Kaskman et al. (2019), and NOCS Wang et al. (2019); and outdoor scenes from nuScenes Caesar et al. (2020). We transform the object annotations into a gravity-aligned frame anchored to the reference camera, using the ground or tabletop as the height reference. Selected views provide complementary observations of the scene, and projection associates the queried 3D boxes with image regions. Each instance pairs these multi-view observations and object queries with oriented-box annotations in the shared frame, requiring the partial observations to be combined into a joint layout.

B.4 NAVIGATION

Offline path prediction. Navigation episodes use first-person images and camera trajectories from ScanNet Dai et al. (2017), with obstacle maps derived from EmbodiedScan object geometry Wang et al. (2024a). We organize the observations into a numbered tour sequence and select a starting viewpoint and a target object. The start is specified by an image frame, with a centred scene object indicating the initial heading; the goal is specified by its object name, frame index, and image bounding box. Its metric coordinates must be inferred from the observations.

The task is to predict a path that avoids furniture and obstacles and reaches the target's vicinity. The output is an ordered sequence of ground-plane waypoints in metres, with the origin at the start, x pointing right, and y forward, serialized as

```
<bev_task_pred coords="0 x1 y1; 0 x2 y2; ...  0 xn yn">
</bev_task_pred>
```

Here, `0` is the fixed track identifier of the moving agent. Reference routes are obtained from recorded camera-trajectory segments or planned on the obstacle map and expressed in the same start-centred frame. Evaluation assesses target arrival and collision avoidance, allowing alternative feasible routes.

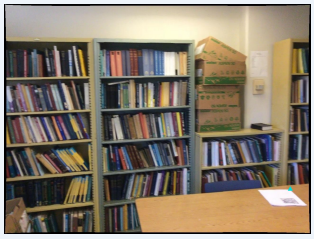
B.5 TASK EXAMPLES

Each example groups its visual input, prompt, and reference answer. The object-correspondence task includes matching and missing-instance variants.

01 Pixel correspondence — Geometric perception

VISUAL INPUT

View 1



View 2



PROMPT

The Cyan triangle in the first image marks a specific location. Which colored marker in the second image shows the same physical point? Choices: A. Lime. B. Magenta. C. White. D. Purple. E. Green. F. Yellow. G. Pink. H. Orange. Please answer with the correct option.

REFERENCE ANSWER

H. Orange

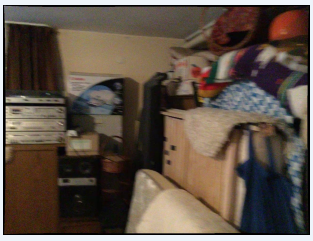
02 Camera pose — Geometric perception

VISUAL INPUT

View 1



View 2



PROMPT

These two images were captured by a moving camera: the first image is viewpoint 1 and the second image is viewpoint 2 of the same scene. Relative to the FIRST camera (looking along its own forward direction), in which direction did the camera MOVE from viewpoint 1 to viewpoint 2? Choices: A. Forward B. Backward-left C. Left D. Forward-left E. Backward F. Right G. Backward-right H. Forward-right Please answer directly with only the letter of the correct option and nothing else.

REFERENCE ANSWER

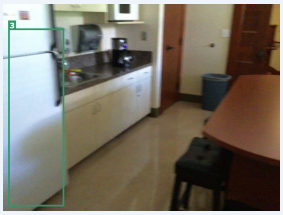
A — Forward

03 Depth ordering

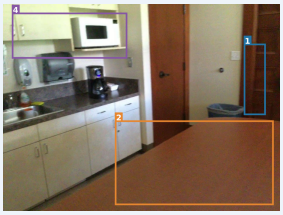
Geometric perception

VISUAL INPUT

View 1



View 2



PROMPT

Two photos show the same room from different viewpoints. Consider these 4 objects: object 1 = the doorframe (box [869, 192, 949, 537] in image 2); object 2 = the table (box [404, 562, 977, 972] in image 2); object 3 = the refrigerator (box [19, 118, 221, 978] in image 1); object 4 = the cabinet (box [32, 42, 449, 264] in image 2). Order them by DEPTH along IMAGE 2's viewing direction, from NEAREST (closest to the image-2 camera) to FARTHEST. Each option lists object numbers nearest-first.

Options: A. (4, 3, 1, 2) B. (3, 2, 4, 1) C. (3, 4, 1, 2) D. (1, 2, 4, 3) E. (1, 2, 3, 4) F. (1, 4, 3, 2) G. (2, 3, 4, 1) H. (4, 3, 2, 1)

Answer with a single option letter.

REFERENCE ANSWER

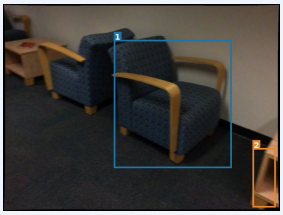
G. (2, 3, 4, 1)

04 Metric depth

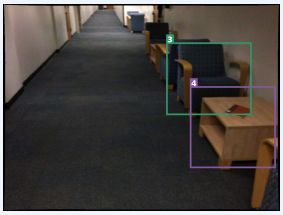
Geometric perception

VISUAL INPUT

View 1



View 2



PROMPT

Two photos show the same room from different viewpoints. Consider these 4 objects: object 1 = the chair (box [400, 173, 826, 793] in image 1); object 2 = the table (box [901, 696, 985, 983] in image 1); object 3 = the chair (box [589, 185, 898, 537] in image 2); object 4 = the table (box [675, 397, 985, 793] in image 2). Boxes are [x0, y0, x1, y1] in pixel coordinates normalized to 0-1000 (origin at the top-left, x rightward, y downward).

For EACH object, estimate its DEPTH along IMAGE 1's viewing direction: how far in metres the object's centre is in front of the IMAGE-1 camera, measured along that camera's viewing direction (i.e. its depth, NOT the straight-line distance to the camera).

Answer with real metric values in metres, to two decimals. Do not answer with ranges or words.

Output ONE JSON object and NOTHING else: {"depth_m": {"1": <number>, "2": <number>, "3": <number>, "4": <number>}}

REFERENCE ANSWER

{"depth_m": {"1": 2.0, "2": 1.25, "3": 4.03, "4": 3.46}}

05 Camera-anchored relation

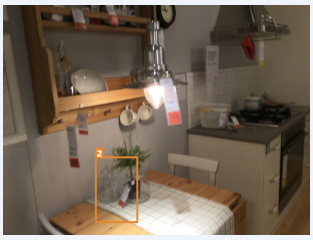
Spatial relations

VISUAL INPUT

View 1



View 2



PROMPT

The two pictures are two camera views of the same room, taken from different spots but sharing some overlapping content. Each object below shows up in only ONE of the pictures and is marked by a bounding box [x0, y0, x1, y1] (normalized 0-1000; origin top-left, x to the right, y downward).

- Object 1 = the oven, box [123, 564, 265, 733] in the FIRST picture.
- Object 2 = the plant, box [300, 659, 440, 948] in the SECOND picture. Put yourself in the shoes of whoever took the FIRST picture, facing the way that picture points. Relative to Object 2, which way is Object 1? (front = further from you into the scene, back = closer to you, left / right = your own left / right.) Options: A. front B. front-right C. right D. back-right E. back F. back-left G. left H. front-left. Reply with a single letter.

REFERENCE ANSWER

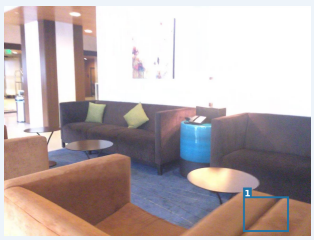
A. front

06 Object-anchored relation

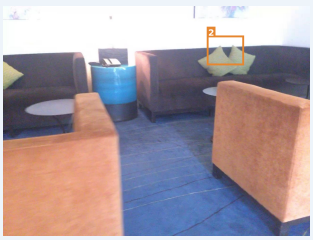
Spatial relations

VISUAL INPUT

View 1



View 2



PROMPT

These two images show the same room from different angles. Object 1 (Image 1): couch, box [779, 828, 928, 982] Object 2 (Image 2): cushion, box [664, 129, 785, 258] Boxes are [x1, y1, x2, y2], normalized to 0-1000. From the perspective of Object 1 (facing its own front), which relative direction is Object 2? Options: A. front | B. front-right | C. right | D. back-right | E. back | F. back-left | G. left | H. front-left. Reply with a single letter.

REFERENCE ANSWER

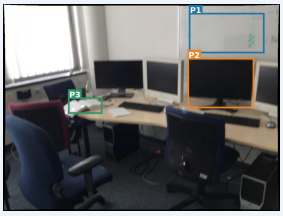
A. front

07 Distance ranking

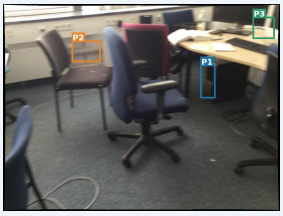
Spatial relations

VISUAL INPUT

View 1



View 2



PROMPT

These two images are different viewpoints of the same room. Boxes are [x0, y0, x1, y1] in pixel coordinates normalized to 0-1000 (origin at the top-left, x rightward, y downward). Consider these three object pairs:

- Pair 1: the blackboard in image 1 (at box [671, 43, 944, 238]) and the computer in image 2 (at box [712, 293, 767, 456])
- Pair 2: the monitor in image 1 (at box [666, 260, 907, 504]) and the pipe in image 2 (at box [248, 176, 353, 281])
- Pair 3: the book in image 1 (at box [235, 451, 359, 530]) and the monitor in image 2 (at box [901, 61, 981, 169]) Rank the pairs from CLOSEST to FARTHEST by the real-world 3D distance between the two objects in each pair (each option lists pair numbers closest-first). Choices: A. (2, 3, 1) B. (1, 2, 3) C. (3, 1, 2) D. (3, 2, 1) E. (2, 1, 3) F. (1, 3, 2). Answer with the option letter only.

REFERENCE ANSWER

C. (3, 1, 2)

08 Nearest object

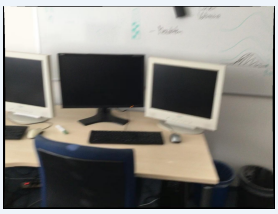
Spatial relations

VISUAL INPUT

View 1



View 2



View 3



PROMPT

These 3 images are different viewpoints of the same room. Boxes are [x0, y0, x1, y1] in pixel coordinates normalized to 0-1000 (origin at the top-left, x rightward, y downward). Reference object: the monitor in image 1 (at box [778, 106, 962, 181]). Candidates:

- A. the desk in image 3 (at box [46, 652, 132, 886])
- B. the blackboard in image 2 (at box [235, 77, 917, 419])
- C. the chair in image 3 (at box [62, 126, 233, 614])
- D. the bin in image 2 (at box [847, 772, 972, 963]) Which candidate is CLOSEST to the monitor in the real 3D room? Each object appears in only one of the images, so you must relate the viewpoints to each other. Answer with the option letter only.

REFERENCE ANSWER

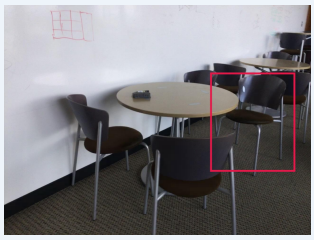
B. the blackboard in image 2

09 Object correspondence: matching

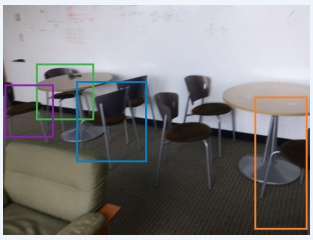
Spatial relations

VISUAL INPUT

Target view



Query view



PROMPT

The first image marks one target chair with a red box. In the last image, several chairs are each marked with a colored box. Which colored chair is the same object as the red-boxed target? Choices: A. green B. orange C. purple D. Not visible E. blue. Please answer directly with only the letter of the correct option and nothing else.

REFERENCE ANSWER

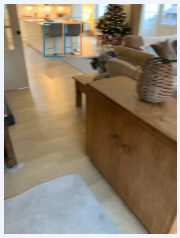
E

10 Object correspondence: elimination

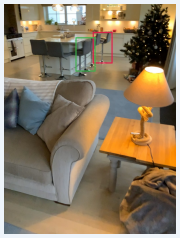
Spatial relations

VISUAL INPUT

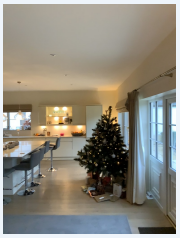
Context 1



Context 2



Query view



PROMPT

The first 2 image(s) mark 4 different bar stools, each with a colored box. The last image shows the same room from another viewpoint. Which colored bar stool is missing from the last image? Choices: A. red B. green C. blue D. orange. Please answer directly with only the letter of the correct option and nothing else.

REFERENCE ANSWER

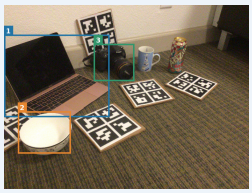
C

11 Joint scene reconstruction

Scene reconstruction

VISUAL INPUT

View 1



View 2



PROMPT

You are given two photographs of the same TABLETOP scene, taken from two nearby viewpoints (both looking down at the table from above) with overlapping fields of view. Reconstruct the 3D layout of the listed objects as a structured scene.
COORDINATE SYSTEM (right-handed, three.js/glTF, units = METERS):
- +Y is up (against gravity); y = 0 is the TABLE surface; an object's y is the height of its box CENTER above the table.
- The horizontal origin (x=0, z=0) is on the table directly below IMAGE 1's camera.
- IMAGE 1 looks DOWN at the table from above. +X is IMAGE 1's right; -Z is the horizontal direction IMAGE 1 faces (its view direction projected onto the table) – NOT its downward line of sight.
OBJECTS TO PLACE (each is marked by a 2D box in ONE image; box is normalized 0-1000, origin top-left, x->right, y->down). Place every listed object:
- id "laptop_1" | laptop | IMAGE 1 | box [0, 162, 431, 610]
- id "bowl_1" | bowl | IMAGE 1 | box [57, 579, 273, 808]
- id "camera_1" | camera | IMAGE 1 | box [367, 212, 532, 414]
- id "mug_1" | mug | IMAGE 2 | box [656, 306, 750, 458]
- id "can_1" | can | IMAGE 2 | box [792, 297, 879, 512]
IMPORTANT — object ids: each 'id' above (e.g. "obj_15_1") is an OPAQUE HANDLE. In your output copy it back EXACTLY, character-for-character, INCLUDING the trailing instance suffix (the "_1"). Do NOT shorten it to the category (e.g. "obj_15"), rename it, or invent new ids. Output exactly one entry per listed id.
FOR EACH listed object output its 3D oriented bounding box:
- center [x, y, z] (meters; y = box-center height above the table)
- size [w, h, d] (meters, in the object's OWN axes: w = width along its local X, h = height along Y/up, d = depth along its local Z/front-back)
- yaw_deg (rotation about +Y; 0 = object's front (+local Z) faces -Z; positive = counter-clockwise seen from above)
GUIDANCE:
- Use the two OVERLAPPING views to triangulate depth and relative position.
- These are SMALL tabletop items; estimate REAL metric sizes from commonsense (typically 3-25 cm). The whole scene must be self-consistent and at real scale.
- Objects rest ON the table (small y) unless clearly stacked on another object.
You may reason first. Then output ONE JSON object and NOTHING after it, using the exact ids given above: "'json {{"objects":[{{"id":"<id>","category":"<cat>","center":[x,y,z],"size":[w,h,d], "yaw_deg":a}}]}} "'

REFERENCE ANSWER

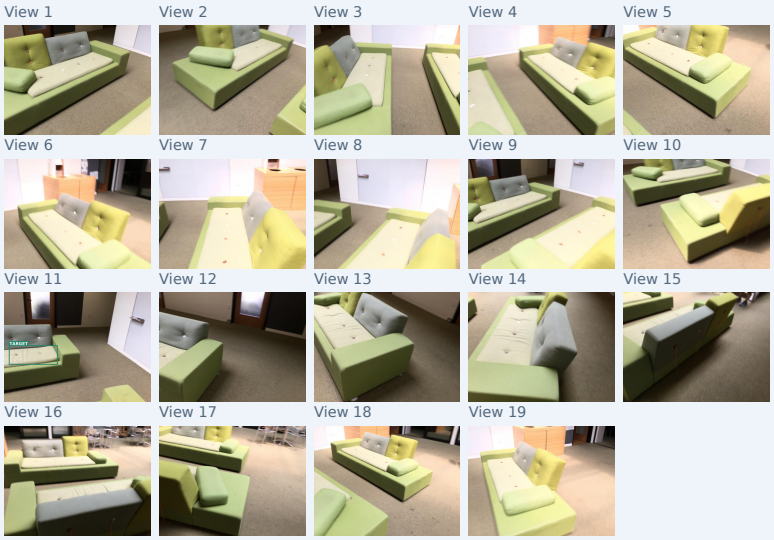
Object	Centre (m)	Extent (m)	Orientation
laptop_1	(-0.252, 0.085, -0.758)	(0.318, 0.169, 0.283)	q=(0.000, -0.881, 0.000, 0.473)
bowl_1	(-0.216, 0.041, -0.477)	(0.144, 0.060, 0.141)	q=(0.000, -0.341, 0.000, 0.940)
camera_1	(-0.036, 0.061, -0.925)	(0.187, 0.105, 0.149)	q=(0.000, -0.970, 0.000, 0.243)
mug_1	(0.115, 0.054, -0.972)	(0.108, 0.103, 0.077)	q=(0.000, -0.784, 0.000, 0.621)
can_1	(0.242, 0.082, -0.979)	(0.070, 0.154, 0.069)	q=(0.000, -0.666, 0.000, 0.746)

Quaternion order: (x, y, z, w). Values rounded to three decimals.

12 Offline navigation

Navigation

VISUAL INPUT



PROMPT

You are given 19 discrete first-person frames (numbered 1-19) of an indoor tour.

Starting Position: You are at the viewpoint of Frame #19, facing the couch centered in the image.

Coordinate System: Origin (0.00, 0.00) is your current position. +y is straight ahead, +x is to your right (in meters). Each waypoint is a coordinate (x, y) in this local frame. Navigation Goal: Walk to the mat shown in Frame #11 (bounding box [33, 484, 365, 662], normalized pixels 0-1000).

No numeric goal coordinate is provided — estimate the distances yourself from the tour images. Report the path in meters as best you can (only its overall direction and shape are scored). Task: Plan a collision-free path to the goal, routing around furniture and obstacles rather than heading straight for it (a direct line is often blocked).

Output Format: Output ONLY the planned path XML tag. Use the prefix 0 (a fixed track id for you, the mover) before each waypoint: <bev_task_pred coords="0 x1 y1; 0 x2 y2; ... 0 xn yn"></bev_task_pred> (Each x, y in meters to 2 decimal places. The first waypoint should be near 0.00 0.00, and the last should be close to the goal.)

REFERENCE ANSWER

<bev_task_pred coords="0 0.00 0.00; 0 -0.38 0.68; 0 -1.06 2.77; 0 -2.10 2.70"></bev_task_pred>

Table 5: Anatomy of the reconstruction scores. All components are deterministic and higher is better. For nondegenerate Structure channels, random-layout correction maps the empirical random reference to zero and a ground-truth copy to one.

Score	Component	Quantity tested
Structure	Adjacency	Jaccard agreement of local neighbor graphs
	Shape	Residual after planar similarity alignment
	Axis order	Chance-corrected Kendall order on three axes
Metric	Position	AP over normalized localization thresholds
	View IoU	Mean top-down and front-view box overlap
	Scale	Penalty for the fitted global scale correction

C Reconstruction Output and Scoring Details

Reconstruction instances query an average of 6.73, 5.64, and 8.62 objects in indoor, tabletop, and outdoor scenes, respectively. Their geometric annotations derive from registered/FARO scans, object poses and extents, and LiDAR boxes.

Coordinate convention. Centres and extents are expressed in metres. The horizontal origin lies below the reference camera on the support plane: the floor or ground for indoor and outdoor scenes, and the table surface for tabletop scenes. The $+x$ axis points right, $+y$ upward, and $-z$ follows the reference camera's horizontal viewing direction; yaw is measured about gravity. The first view is the reference, using the front camera for nuScenes. Camera intrinsics, full poses, and depth maps are not provided.

Table 5 summarizes the reconstruction components.

Parsing and matching. The required response is a JSON object with an `objects` list. Each entry contains `id`, `category`, `center`, `size`, and `yaw_deg`; angles are converted to radians internally. The parser first selects the last fenced JSON object, or otherwise searches the response for a decodable object containing the required list. It performs no generative syntax repair. Entries without an identifier or with fields that cannot be converted to the required numerical arrays are discarded. Missing or empty centre and size arrays default to $(0, 0, 0)$ and $(0.4, 0.4, 0.4)$ metres; missing yaw defaults to zero and missing category to `object`. Predictions are matched only to queried object IDs, keeping the first parsed occurrence of a duplicate ID. Fewer than three matches gives zero on both headline scores. Otherwise, geometric components use the matched subset and coverage is reported separately. Empty and unparseable scenes remain in the headline average; geometric diagnostics are averaged only where defined. Category-label F1 compares the multisets of supplied and returned labels, measuring label consistency rather than independent recognition. Returned categories do not affect ID matching, and neither category F1 nor coverage enters the two reconstruction scores.

Model references. Model references point to version-specific releases where available and to family-level technical reports otherwise. The explicit version names in Section 4 identify the evaluated systems. The Gemini-3.1-Pro label denotes the `gemini-3.1-pro-preview` endpoint.

Saved-run coverage and prompt versions. For the full 840-instance evaluation, a record is identified by its source batch and `gt_key`; repeated attempts within one batch retain the last saved response. All 18 models have results in all three domains, with 836–840 saved reconstruction instances per model. The 19 absent model–instance records (18 indoor and one outdoor) are excluded from each model's denominator; saved empty or unparseable outputs remain as errors. No common-instance intersection is imposed on the leaderboard. The full-precision public result is the source for all manuscript headline tables and figures.

Historical tabletop runs of Claude-Opus-4.7, Claude-Opus-4.8, Gemini-3.5-Flash, and Qwen3.7-Plus each contain 89 prompts without a later explicit reminder to copy the object identifiers character-for-character; 85 Gemini-3.1-Pro prompts use that earlier wording. The queried objects, their image regions, and the remaining task text agree. We retain these saved runs and apply the same fixed parser

and scorer; rescoring does not remove the historical prompt difference. GPT-6-Astra was evaluated through the uniform-eval integration, while the other models use the benchmark's model-specific API runners.

Gemini-3.8-Flash and Gemini-3.7-Flash each cover all 840 reconstruction instances. On navigation, Gemini-3.8-Flash has 59 unparseable responses among 346 episodes, including 57 that place the `coords` field inside the prediction tag's body rather than in its required attribute. These responses score zero under the unchanged parser; its reported 12.1% success therefore includes output-format failures. Gemini-3.7-Flash likewise retains nine unparseable responses in its denominator. No model-specific parser relaxation is applied.

Numerical reproducibility. Each instance uses a fresh random stream initialized from the first eight bytes of SHA256(`gt_key—20260911`), interpreted as a big-endian integer. The published snapshot fixes the scorer and its numerical environment (NumPy 2.5.1 and SciPy 1.17.1); the release includes file hashes and per-instance scores. Diagnostic components used in the cross-level analysis were recomputed in that environment, with domain means verified against the official snapshot for all eighteen models.

Raw Structure components. Let $g_i, p_i \in \mathbb{R}^2$ be matched ground-truth and predicted centres in the horizontal (x, z) plane. For their undirected k-nearest-neighbor edge sets E_G, E_P, with $k = \min(3, n-1)$, the raw similarities are

$$r_{\text{adj}} = \frac{|E_G \cap E_P|}{|E_G \cup E_P|}, \tag{5}$$

$$r_{\text{shape}} = \max\left\{0, 1 - \frac{\sum_{i=1}^{n} \|g_i - \phi(p_i)\|_2}{n(D/2+\epsilon)}\right\}, \tag{6}$$

where D is the diagonal of the ground-truth centres' axis-aligned bounding box and $\phi(p) = aRp+t$ is the least-squares planar similarity fit with $R \in \mathrm{SO}(2)$. The residual uses the fixed ID correspondences rather than nearest-point matching. For each coordinate axis, the raw order score is Kendall's tie-adjusted correlation

$$r_{\text{axis}} = \tau_b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_G)(N_c + N_d + T_P)}}, \tag{7}$$

where N_c, N_d count concordant and discordant object pairs, and T_G, T_P count pairs tied only in the ground-truth or predicted ordering. Constant or undefined orderings receive raw score zero.

Random-layout calibration. Each raw similarity r_j is calibrated against a reference b_j:

$$S_j = \mathrm{clip}_{[0,1]}\left(\frac{r_j - b_j}{1 - b_j + \epsilon}\right), \qquad \epsilon = 1/10^9. \tag{8}$$

Horizontal references sample n points uniformly within the ground-truth bounding box; vertical references sample n heights within its height range. Each reference averages 20 draws. This measures agreement beyond the empirical random baseline. An uninformative component with $r_j = b_j = 1$, such as a necessarily complete neighbor graph, receives zero after calibration.

Metric alignment. For matched 3D centres G_i, P_i, the scale-only least-squares fit is

$$s = \frac{\sum_i \langle G_i, P_i \rangle}{\sum_i \|P_i\|_2^2 + \epsilon}. \tag{9}$$

No rotation or translation is fitted. The position component averages the fraction of matched objects with $\|G_i - sP_i\|_2 < tR_g$ over $t \in \{0.1, 0.2, 0.4, 0.8\}$, where $R_g = \sqrt{\frac{1}{n}\sum_i \|G_i - \bar{G}\|_2^2 + \epsilon}$ is the ground-truth radius of gyration with numerical stabilization. View IoU averages per-object polygon overlap in the horizontal (x, z) and front (x, y) projections, scaling both predicted centres and box extents by s. The separate scale score is $\max\{0, 1 - |\log \max(s, 1/10^6)|/\log 2\}$. Absolute position, size, and yaw errors are retained as diagnostics.

Table 6: Reconstruction scorer controls on the original 457-instance subset, with instance averaging within each of the three domains and equal domain weights. The shuffled+scale condition supplies a joint empirical floor; depth-flattened is an idealized ablation that retains all other ground-truth quantities.

Control	Structure	Metric	Pos.	View IoU	Scale
GT copy	100.0	100.0	100.0	100.0	100.0
Shuffled boxes	9.2	23.3	11.6	7.3	51.1
Depth flattened	52.1	65.0	36.7	58.3	100.0
Random scale	100.0	77.0	100.0	100.0	30.9
Shuffled + scale	9.2	13.8	12.5	8.7	20.0

C.1 Scorer controls

The controls in Table 6 isolate different geometric properties on the fixed 457-instance diagnostic subset. A GT copy establishes the upper reference, shuffling changes object assignments, random scaling changes global calibration, and depth flattening selectively corrupts one coordinate while retaining the remaining GT information. The combined shuffled-and-scaled condition provides an empirical joint floor. Each condition uses fixed random seeds and the same matching definitions as model evaluation; instances are pooled within indoor, tabletop, and outdoor before the three domains are averaged equally. In particular, depth flattening is an idealized corruption experiment, not a performance ceiling for image-based prediction.

D Per-Domain and Cross-Level Analysis

The domain decomposition in Table 7 retains the physical regimes obscured by a single reconstruction average. Structure exceeds Metric for every model in every domain, while the magnitude of the disparity varies with scene scale and capture geometry. ScanNet and CA-1M contribute to one indoor average, computed directly over their available instances.

Uncertainty over repeated scene configurations. The reconstruction intervals use 10,000 bootstrap replicates, stratified by reporting domain. Clusters are source-qualified annotation scene identifiers; for outdoor batches these identify keyframes shared by multiple query configurations. All instances of a sampled identifier receive the same cluster weight. Domain means remain instance-weighted within each replicate, and the three domains receive equal weight. This accounts for configurations sharing an identifier without assuming that every released instance is independent. It does not eliminate all possible temporal or source-level dependence. The 95% intervals for the three-domain Structure–Metric gap exclude zero for all eighteen models. GPT-6-Astra's gap is 0.226, with interval $[0.206, 0.245]$. These intervals concern the aggregate gap; the consistently positive gaps in all 54 model–domain combinations are reported descriptively in Table 7.

Residual cross-level associations. At the level of aggregate task scores, Geometric Perception correlates with Structure ($r = 0.83$) and Metric ($r = 0.86$); the corresponding Spatial Relations correlations are 0.69 and 0.81. These raw associations combine shared differences in model strength with any task-specific relationship. For each atomic score x and reconstruction component y, let z be the mean of the other eight atomic scores across the same eighteen models. We regress x and y separately on an intercept and z, then correlate the residuals:

$$r_{\text{partial}} = \operatorname{corr}\big(x - \widehat{x}(z),\ y - \widehat{y}(z)\big). \tag{10}$$

Figure 6 reports this matrix. Headline scores retain failed predictions as zero; component diagnostics are averaged over instances where the component is defined. Leave-one-model-out ranges refit the analysis after each model is removed. They quantify sensitivity to the evaluated cohort, not confidence intervals for a population of models. The associations remain observational; causal information effects are assessed through the matched prompt interventions.

Table 7: Structure (S) and Metric (M) reconstruction scores on the full 840-instance release. ScanNet and CA-1M are pooled into Indoor; N is each model's actual saved-instance count. Values are percentages.

Model	Indoor			Tabletop			Outdoor		
	N	S	M	N	S	M	N	S	M
GPT-6-Astra	475	72.5	50.9	89	87.4	55.8	276	84.1	69.4
GPT-5.6-Sol	473	56.8	34.9	89	82.2	41.1	276	74.8	48.4
Claude-Fable-5	475	54.8	29.8	89	74.5	39.4	276	76.8	49.1
Claude-Opus-5	473	53.9	28.4	89	76.8	38.3	276	73.7	50.0
GLM-5.3-Flash	474	56.5	32.6	89	74.6	34.8	275	73.9	44.3
Qwen3.8-Max	474	56.3	34.2	89	77.0	28.5	276	72.1	47.4
Gemini-3.8-Flash	475	55.0	33.5	89	74.7	34.8	276	69.4	44.6
Kimi-K3	474	54.9	33.6	89	77.9	19.8	276	73.7	44.6
Doubao-Seed-2.1-Pro	475	54.9	33.8	89	76.4	24.6	276	59.9	32.3
GPT-5.5	475	54.4	32.1	89	73.6	22.5	276	67.0	31.2
Grok-4.6	475	50.5	30.4	89	71.6	19.2	276	67.9	38.6
Gemini-3.1-Pro	475	53.8	30.1	89	69.6	20.4	276	63.5	38.1
Gemini-3.7-Flash	475	45.1	28.9	89	58.2	28.6	276	57.7	43.3
Gemini-3.5-Flash	475	36.2	25.4	89	58.1	37.5	276	53.0	35.9
Claude-Opus-4.8	471	40.8	19.2	89	60.0	16.9	276	65.6	36.7
Claude-Opus-4.7	471	41.4	22.0	89	61.5	10.4	276	58.7	30.8
Qwen3.7-Plus	472	43.2	15.6	89	72.3	14.1	276	42.6	29.1
MiniMax-M3	475	38.6	20.9	89	54.8	12.8	276	58.9	20.8

Table 8: Cross-level associations over eighteen models. Partial correlation controls for the mean of the other eight atomic units; LOO gives the range after leaving out one model at a time. Raw correlations largely track general model strength, whereas only selected residual associations are stable. Task abbreviations follow Table 1.

Atomic unit	Reconstruction axis	Pearson	Spearman	Partial	LOO range
PC	metric	+0.86	+0.81	+0.46	[+0.38, +0.53]
PC	posAP	+0.78	+0.77	+0.34	[+0.24, +0.42]
CP	vIoU	+0.88	+0.80	+0.62	[+0.29, +0.72]
DO	FB	+0.78	+0.75	+0.63	[+0.51, +0.75]
MD	metric	+0.53	+0.43	+0.04	[-0.10, +0.09]
Rel	LR	+0.63	+0.58	-0.04	[-0.19, +0.07]
Rel	posAP	+0.79	+0.86	+0.34	[+0.23, +0.49]
Dist	structure	+0.60	+0.58	-0.06	[-0.18, +0.09]
Near	posAP	+0.82	+0.79	+0.45	[+0.32, +0.56]

E Diagnostic and Robustness Analyses

E.1 Verification and revision

Figure 7 distinguishes detecting a corrupted layout, identifying missing objects, and improving a scene through revision. The four-model detection probes assess which discrepancies are perceptually accessible, whereas the GPT-5.6-Sol revision comparison measures whether a rendering provides useful corrective information beyond an additional textual turn. The discrepancy between detection and revision is consistent with a representation whose salient arrangement is easier to verify than its metric calibration or completeness.

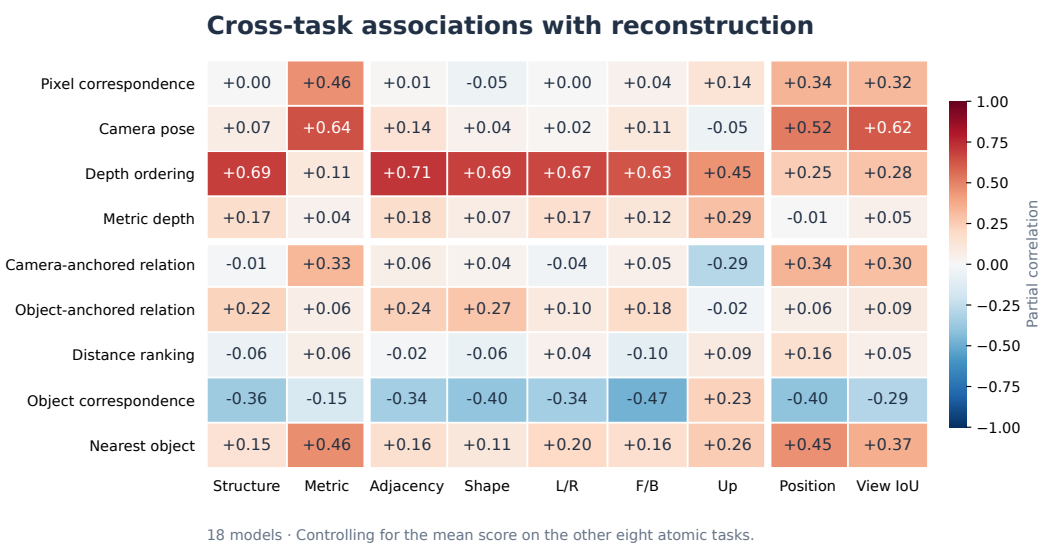


Figure 6: Full cross-level partial-correlation matrix over eighteen models, using the expanded reconstruction suite and three-domain aggregation. Each cell controls for the mean of the other eight atomic units. The matrix reveals a strong localization/pose cluster and a separate ordinal/structure cluster, while several semantically intuitive one-to-one pairings remain weak. Correlations are diagnostic rather than causal because the model sample is small; the matched prompt interventions provide the causal evidence.

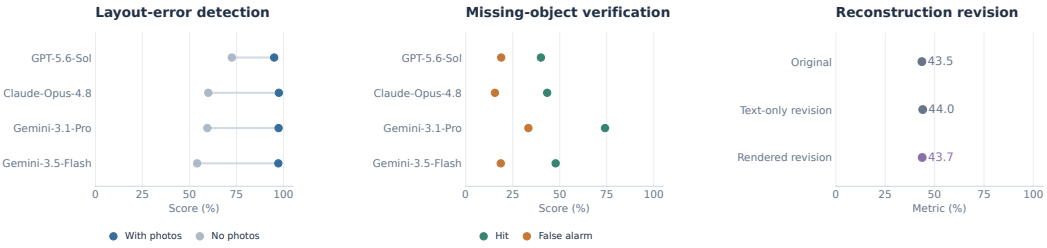


Figure 7: Three complementary properties of an explicit scene. Four models detect large layout corruptions reliably when photographs are supplied, but missing-object verification remains substantially weaker. For GPT-5.6-Sol, rendering its prediction before a second-turn revision does not improve the current Metric score beyond a text-only second turn.

Across 157 instances, adding a rendering of the predicted scene to a second turn changes Metric by only -0.003 relative to a matched text-only revision. Four models detect large layout corruptions more reliably than missing objects. Detecting salient relational discrepancies therefore need not provide an equally useful signal for correcting completeness or metric calibration.

E.2 Incremental Reconstruction

Incremental conditions separate interaction structure from retained evidence. The five-arm comparison uses 40 shared instances, while the three full-history conditions additionally admit a 53-instance comparison (Table 9). Accumulating observations or retaining full visual history yields paired intervals spanning zero relative to the one-shot control. Removing historical images produces decreases of 0.137 in Structure (95% interval $[0.079, 0.196]$) and 0.098 in Metric ($[0.028, 0.164]$), despite retaining every object identifier in the textual map. Stability with full visual history therefore does not establish that the exported map is a sufficient substitute for past observations. These contrasts concern the sufficiency of the retained representation under the tested update protocol, rather than a general assessment of long-term spatial memory.

Table 9: Incremental reconstruction controls for GPT-5.6-Sol, rescored with the current scorer. The first block uses the matched 40-item subset; the second uses all 53 constructible episodes for the available arms.

Subset	Condition	N	Structure	Metric	Position error (m)
Matched	One shot	40	0.581	0.344	1.49
Matched	History images removed	40	0.444	0.246	1.91
Matched	Images accumulated	40	0.593	0.305	1.57
Matched	Full history retained	40	0.593	0.327	1.49
Matched	Matched three rounds	40	0.627	0.327	1.53
Full supply	One shot	53	0.592	0.338	1.52
Full supply	Full history retained	53	0.600	0.319	1.54
Full supply	Matched three rounds	53	0.631	0.315	1.58

Table 10: Spearman correlation between the eighteen-model ranking under each alternative scoring convention and the corresponding Structure or Metric ranking on the expanded reconstruction suite.

Alternative scoring convention	Rank Spearman vs. current
equal structure	1.000
metric pos heavy	1.000
metric scale heavy	0.994
metric without scale	0.953
scale tolerance 1.5	0.996
scale tolerance 2	1.000
scale tolerance 3	0.988

E.3 Sensitivity to scoring choices

Alternative scores are computed from the same per-instance components of the 840-instance evaluation, preserving failed outputs as zero and equal weights across the three domains. Table 10 compares the resulting model rankings with the current ranking. Positive reweightings of Structure or Metric components and changes in scale tolerance retain rank correlations of 0.988–1.000 with the corresponding headline ranking. Removing the scale term yields $\rho = 0.953$ and has the largest effect because it changes what the metric measure requires: relative geometry corrected by rescaling no longer incurs a separate calibration penalty. Sensitivity is therefore interpreted in relation to the property removed, rather than solely as a numerical perturbation of the aggregate.

The position and overlap components already allow one fitted scale, while the scale term separately measures its departure from physical calibration. Residual error therefore includes disagreement in scene configuration beyond a uniform change of units.

F Navigation Evaluation Protocol

A predicted path succeeds if its endpoint is within 1 metre of the target footprint and the path satisfies the collision criterion with 0.2-metre lateral clearance tolerance. The target object is excluded from collision checking. Evaluation is route-agnostic: a valid alternative path need not coincide with the reference traversal. The same episode set, path parser, and geometric criteria are used across the map-substitution conditions.

Scale conventions. The navigation leaderboard follows the scale-normalized S2 protocol. For a nonzero predicted endpoint, one scalar aligns its distance from the origin with that of the reference endpoint:

$$s_{\mathrm{nav}} = \|\mathbf{p}_T^{\mathrm{GT}}\|_2 / \|\widehat{\mathbf{p}}_T\|_2, \qquad \mathbf{q}_t = s_{\mathrm{nav}} \widehat{\mathbf{p}}_t. \tag{11}$$

Reach and collision are then evaluated on the transformed path $\{\mathbf{q}_t\}$. The map-substitution study instead evaluates unscaled metric paths, so a supplied metric map must be used without this correction.

Table 11: Detailed 3D box map results on 346 offline navigation episodes per model, using unscaled metric paths. Tier 1 uses human traversals; Tier 2 requires algorithmically generated detours. Human/solver reference success is 97.7%; a straight-to-goal path succeeds on 14.2%. Scores are percentages; each model serves as its own planner in all four conditions.

Model	Input	Success	Reach	Collision-free	Tier 1	Tier 2
GPT-5.6-Sol	Images	15.6	41.3	27.7	18.7	11.8
	Self-map	15.6	43.1	27.7	15.5	15.7
	GT map	85.3	92.5	92.5	82.4	88.9
	GT + target ID	90.5	100.0	90.5	85.5	96.7
GPT-6-Astra	Images	22.5	61.8	29.2	21.8	23.5
	Self-map	35.5	66.2	49.4	37.8	32.7
	GT map	97.1	99.4	97.4	95.3	99.3
	GT + target ID	97.4	99.7	97.7	95.3	100.0
Qwen3.8-Max	Images	9.5	31.2	16.2	9.8	9.2
	Self-map	11.0	34.1	21.4	12.4	9.2
	GT map	66.2	94.2	68.8	59.1	75.2
	GT + target ID	78.6	100.0	78.6	75.1	83.0
Gemini-3.8-Flash	Images	11.8	38.4	21.7	11.9	11.8
	Self-map	18.2	43.9	28.6	18.1	18.3
	GT map	82.1	94.8	85.3	77.7	87.6
	GT + target ID	79.8	99.1	80.1	71.5	90.2

Accordingly, its image-only reference for GPT-5.6-Sol, for example, is 0.156 rather than the 0.217 navigation leaderboard score. Neither protocol executes actions in an interactive environment.

3D box map representation. The episode frame has its origin at the starting position, $+x$ to the right, $+y$ forward, and $+z$ upward. This differs from the reconstruction task's y-up convention. Maps use the same object-centric parameterization as the reconstruction task: object IDs, categories, centres, dimensions, and yaw, with yaw measured counterclockwise from $+x$ to the width axis. The self-map condition first asks the model to select navigation-relevant objects and predict their boxes from the episode observations, then supplies those boxes for path prediction. These are episode-specific predictions, not outputs reused from the reconstruction leaderboard. GT object boxes are derived from EmbodiedScan (Wang et al., 2024a) annotations in the episode frame. Synthetic boundary boxes have a nominal height of 2.5 metres; collision evaluation uses ground-plane geometry.

Information varied by the intervention. The GT 3D box map supplies geometry while leaving the model to associate the visual target with a map entry; an additional condition explicitly identifies that entry. The self-map condition reuses the model's own target association when available. Prompts describe GT maps as exact and self-generated maps as approximate. The comparisons therefore measure the effect of supplied scene information, including its provenance and target association, under a fixed planner and evaluation protocol.

The comparison includes GPT-5.6-Sol, GPT-6-Astra, Qwen3.8-Max, and Gemini-3.8-Flash. Each condition retains all 346 episodes in the evaluation denominator. Table 11 reports outcome components and episode tiers, and Table 12 reports within-model paired effects. Intervals use 10,000 paired episode resamples. Since historical inference configurations vary across models, the intervention analysis relies on within-model contrasts.

GT maps improve success in both episode tiers for all four models and outperform self-generated maps by 55.2–69.7 percentage points, with every paired interval above zero. Self-map gains are reliable for GPT-6-Astra and Gemini-3.8-Flash, but not statistically distinguishable from zero for Sol or Qwen. Target identification adds 5.2 points for Sol and 12.4 for Qwen; its intervals include zero for Astra and Gemini. These contrasts support the utility of accurate scene information for fixed planners, with model-dependent limits in both map construction and target grounding.

Table 12: Paired navigation gains in percentage points with 95% bootstrap intervals over 346 episodes. Target ID gain compares GT map + target ID against GT map. Differences are computed before rounding.

Model	Self-map − Images	GT map − Images	GT map − Self-map	Target ID gain
GPT-5.6-Sol	+0.0 [-4.6, +4.6]	+69.7 [+64.5, +74.9]	+69.7 [+64.5, +74.6]	+5.2 [+1.4, +9.2]
GPT-6-Astra	+13.0 [+7.5, +18.5]	+74.6 [+69.9, +79.2]	+61.6 [+56.4, +66.8]	+0.3 [-0.6, +1.4]
Qwen3.8-Max	+1.4 [-2.3, +5.2]	+56.6 [+51.2, +62.1]	+55.2 [+49.4, +61.0]	+12.4 [+7.5, +17.6]
Gemini-3.8-Flash	+6.4 [+2.0, +10.7]	+70.2 [+65.0, +75.1]	+63.9 [+58.7, +69.1]	-2.3 [-6.6, +2.0]

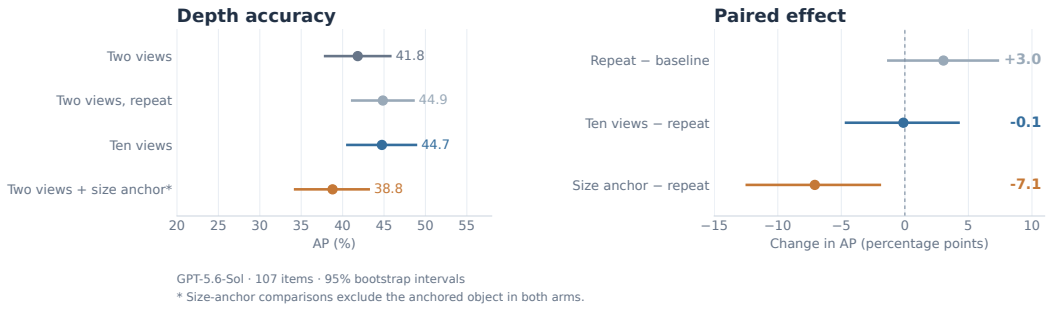


Figure 8: Matched metric-depth controls for GPT-5.6-Sol. Error bars on the left are 95% bootstrap intervals; the right panel reports paired changes. The estimated effect of ten views is not distinguishable from zero relative to the matched two-view repeat, while a correct size anchor does not improve the remaining depths.

G Controlled Information Interventions

Additional views and a size anchor. The metric-depth study on GPT-5.6-Sol uses 107 items with fixed requested depths, parser, and scorer. The two-view repeat is byte-identical to the original prompt and provides a reference for response variability. The additional-view condition retains the two designated images and appends eight unmarked views, with every answer still expressed along the original reference-camera axis. The size-anchor condition supplies one queried object's true longest dimension, excludes that object from the requested predictions, and compares the same remaining three objects with the repeat condition. AP averages correctness at relative-error thresholds $\{5, 10, 20, 40\}\%$; paired percentile-bootstrap intervals use 10,000 item resamples.

Performance is limited both for objects marked in the reference view (AP 0.454) and for those marked in another view (0.376), indicating that the difficulty is not confined to cross-view objects. Adding eight views to the two-view repeat changes AP by -0.001, with a 95% interval of $[-0.047, +0.043]$. Supplying the true size of one object likewise fails to improve the remaining depths (Figure 8). Neither increased visual coverage nor a single scale anchor reliably supplies the missing precision under these controls.

Injecting spatial facts. The fact-injection study uses the same 159 ScanNet instances for GPT-5.6-Sol, GPT-6-Astra, Qwen3.8-Max, and Gemini-3.8-Flash. Each model receives numerical depth and ordinal facts, paired with their respective shuffled controls. Table 13 reports both score changes; Table 14 directly compares their Metric effects. All four intervention arms use all 159 instances in each model.

Exact-pose and depth-plus-pose conditions remain additional probes on GPT-5.6-Sol, compared with the shuffled-depth condition. Depth specifies each query object's signed view-axis position in the reference frame, and pose specifies the second camera's translation and yaw. Historical Base predictions cover 157 Sol, 159 Astra, and 158 Qwen instances; Base is not used in the primary intervention-to-control contrasts, which each use all 159. Structure random references are seeded per instance, and paired intervals are computed from 10,000 paired task-instance bootstrap resamples.

Table 13: Fact-injection effects relative to each condition's shuffled control. Values are percentage-point changes with paired 95% bootstrap intervals; each comparison uses 159 instances.

Model	Supplied information	ΔStructure [95% CI]	ΔMetric [95% CI]
GPT-5.6-Sol	Numerical depth	+23.6 [+20.3, +26.9]	+15.6 [+13.7, +17.4]
	Ordinal facts	+21.6 [+18.4, +24.9]	-1.9 [-4.2, +0.4]
GPT-6-Astra	Numerical depth	+9.4 [+7.1, +11.7]	+8.8 [+7.2, +10.4]
	Ordinal facts	+13.6 [+10.8, +16.5]	-0.8 [-2.9, +1.3]
Qwen3.8-Max	Numerical depth	+19.0 [+15.4, +22.5]	+12.6 [+11.0, +14.4]
	Ordinal facts	+13.2 [+9.8, +16.7]	+2.4 [+0.1, +4.7]
Gemini-3.8-Flash	Numerical depth	+18.4 [+14.5, +22.3]	+11.5 [+9.4, +13.5]
	Ordinal facts	+22.0 [+18.3, +25.9]	+4.6 [+2.1, +7.1]

Table 14: Difference between numerical-depth and ordinal-fact Metric gains, paired over the same 159 instances. Values are percentage points with 95% bootstrap intervals. All four intervals exclude zero.

Model	Numerical − ordinal ΔMetric [95% CI]
GPT-5.6-Sol	+17.5 [+14.6, +20.3]
GPT-6-Astra	+9.5 [+7.0, +12.2]
Qwen3.8-Max	+10.3 [+7.6, +12.9]
Gemini-3.8-Flash	+6.9 [+3.9, +9.9]

True-fact prompts describe the supplied statements as measured and true, whereas shuffled prompts describe them as automatically generated and potentially unreliable. These contrasts therefore include both fact content and its stated reliability, rather than isolating content alone.

In the additional Sol probes, providing both depth and pose leaves substantial residual error: lateral placement, extent, orientation, and cross-object consistency remain to be inferred. The interventions identify useful information rather than a unique latent cause of all reconstruction errors, supporting numerical geometry as a major limitation without reducing reconstruction to a single depth estimate.

H Training and Evaluation Settings

Spatial supervision combines Geometric Perception, Spatial Relations, and scene reconstruction to connect local judgments with consistent 3D scene geometry. We fine-tune Qwen3.6-Flash on 50k examples using reasoning traces and answers generated by Qwen3.8-Max, without general training data. We follow source-level training/evaluation splits. Optimization uses 98 steps, a global batch size of 512, a learning rate of 7×10^{-6}, and 5% warmup. All variants share decoding settings (temperature 0.6 and top-p 0.95) for comparison on the external benchmarks listed in Table 15. For ReconSpatial, the 10k, 20k, and 50k spatial variants share the same data recipe and form a dose series that varies the training budget. All evaluated variants use a common local inference pipeline. The starting checkpoint retains its original long-context positional encoding and serves as a reference.

Table 15: External evaluation benchmarks and the header names used in Table 3.

Table header	Benchmark and evaluated setting
Spatial benchmarks	
SITE	SITE-Bench (Wang et al., 2025)
BLINK	BLINK (Fu et al., 2024)
ViewSpatial	ViewSpatial-Bench (Li et al., 2025a)
ERQA	ERQA (Gemini Robotics Team, 2025)
OmniSpatial	OmniSpatial (Jia et al., 2026)
3DSR	3DSRBench (Ma et al., 2025)
MMSI	MMSI-Bench (Yang et al., 2026a)
EmbSpatial	EmbSpatialBench (Du et al., 2024)
SITE Video	SITE-Bench-Video (Wang et al., 2025)
RefSpatial	RefSpatialBench (Zhou et al., 2025)
Multimodal knowledge	
MMMU	MMMU (val) (Yue et al., 2024)
MMBench CN	MMBench-CN (v1.1) (Liu et al., 2024)
Visual reasoning	
VisualPuzzles	VisualPuzzles (direct) (Song et al., 2025)
BabyVision	BabyVision (Chen et al., 2026)
Chart and diagram understanding	
CharXiv	CharXiv (reasoning) (Wang et al., 2024b)
AI2D	AI2D (Kembhavi et al., 2016)
Scene understanding and perceptual reliability	
RealWorldQA	RealWorldQA (xAI, 2024)
HallusionBench	HallusionBench (Guan et al., 2024)

I Qualitative Case Studies and Provenance

The qualitative analysis examines how relative arrangement, metric anchoring, and representation use can diverge in individual examples. Reconstruction cases hold the input and display convention fixed across four models; the navigation case varies supplied information for the same planner; atomic cases expose specific disagreements in local judgments. These are purposively selected illustrations, including failures of the aggregate leader, rather than a sample for estimating error prevalence. Case R1 appears in the main text and the remaining visualizations are provided here. Query colors and boxes follow the prompt annotations, while all model scenes retain their raw metric coordinates and common axis limits.

Table 16: Case identifiers. Reconstruction IDs combine the batch name and query key, preserving distinct configurations of the same underlying scene.

Case	Saved instance identifier
R1	`scannet8f/scene0628_02_tour8`
R2	`tabletop/nocs_scene_1_0087_0359`
R3	`nusc_c_far8/3a064c48e47b489a8b442b8081f877b6`
R4	`tabletop/tless_000019_1_115`
N1	`scene0148_00_ns_t2_10_01090_a1_s2`
A1	`scene0047_00_00110_01310_depth2`
A2	`scene0542_00_rel_00110_18_29`
A3	`scene0118_02_00090_00170_depthm`

The case bundle preserves inputs, final responses, GT geometry, parsed predictions, scores, and source hashes. Reconstruction scores match the 840-instance result snapshot; navigation follows the unscaled intervention protocol; and atomic outputs use the benchmark parsers. Visual annotations identify the queried objects without changing the evaluated inputs.

The saved R1 prompt retains the introductory wording "two photographs" despite containing eight image attachments and query regions referring to those views. All four models receive that same saved input. We preserve this historical wording in the case bundle and interpret R1 as a comparison of outputs under that configuration, not as a controlled ablation of view count.

Physical calibration remains scene-dependent. Claude-Opus-5 produces the most accurate metric reconstruction in R2 despite GPT-6-Astra's higher aggregate score (Figure 9). In R4, GPT-6-Astra preserves relative arrangement but places the object group too far from the camera, obtaining 85.8 on Structure and 3.8 on Metric (Figure 11). These local contrasts expose errors obscured by either a single scene-quality score or a model's overall rank.

Relative agreement can conceal coordinate error. The far-range outdoor example extends the R1 observation to a different physical scale (R3, Figure 10): plausible barrier ordering coexists with several metres of displacement. Complete object coverage and relational regularity do not sufficiently constrain the anchored coordinates for metric use.

Accurate 3D box maps support collision-free planning. The navigation case distinguishes geometric proximity from a valid decision (Figure 12). In N1, the path generated from the self-generated 3D box map terminates inside the target footprint but intersects obstacles; the GT 3D box map supports a path satisfying both arrival and collision constraints. The contrast illustrates why an explicit scene representation must preserve obstacle geometry to support safe arrival.

Atomic diagnostics localize different failures. Figure 13 illustrates complementary errors in ordering, direction, and numerical depth. Model rankings vary across these local judgments, and a correct directional category does not determine a distance in metres. In A3, the chair marked in the second view lies 4.03 m along the first view's depth axis, while predictions range from 1.60 to 2.94 m. These distinct items illustrate what atomic measurements reveal without constituting a paired causal test of joint reconstruction.

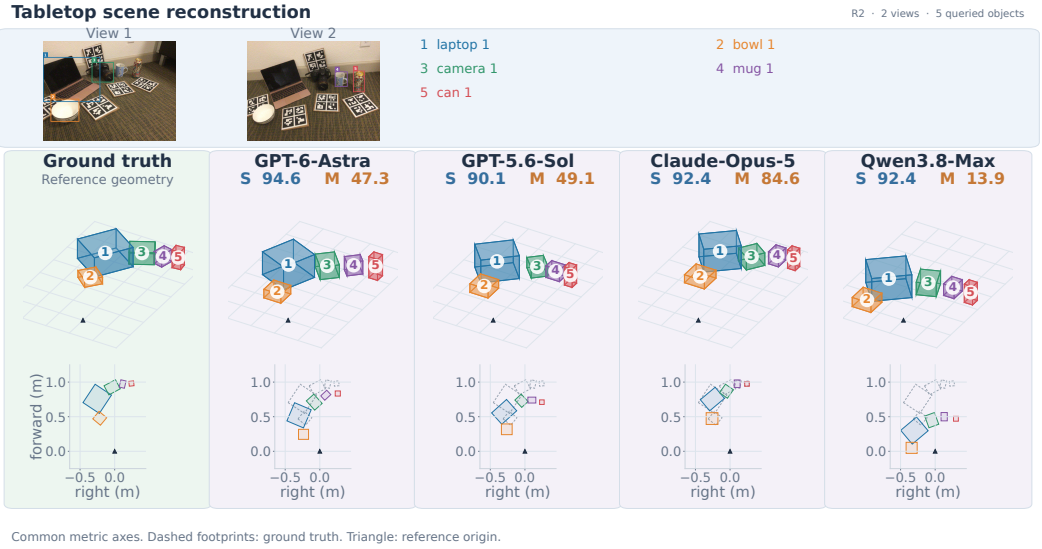


Figure 9: Tabletop case R2. Claude-Opus-5 has the highest Metric score in this local comparison despite a lower aggregate reconstruction score than GPT-6-Astra. All four models retain strong relative structure. The two raw input images and five queried objects are shown above the 3D and ground-plane views. The footprint overlays reveal metric discrepancies that are less apparent from layout alone.

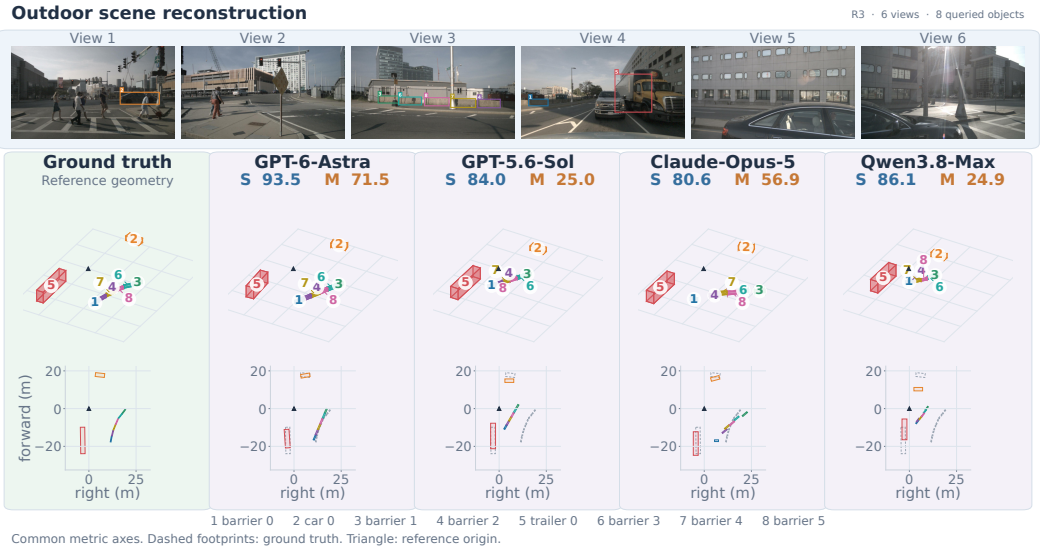


Figure 10: Far-range outdoor case R3. All six surround-camera inputs are shown. The repeated barriers retain a recognizable spatial arrangement across models, while absolute positions and scale differ. All panels share the same metric coordinate limits; the footprints are not rescaled to fill each panel. The ground-plane forward coordinate is the negative of the benchmark's camera-frame z coordinate.

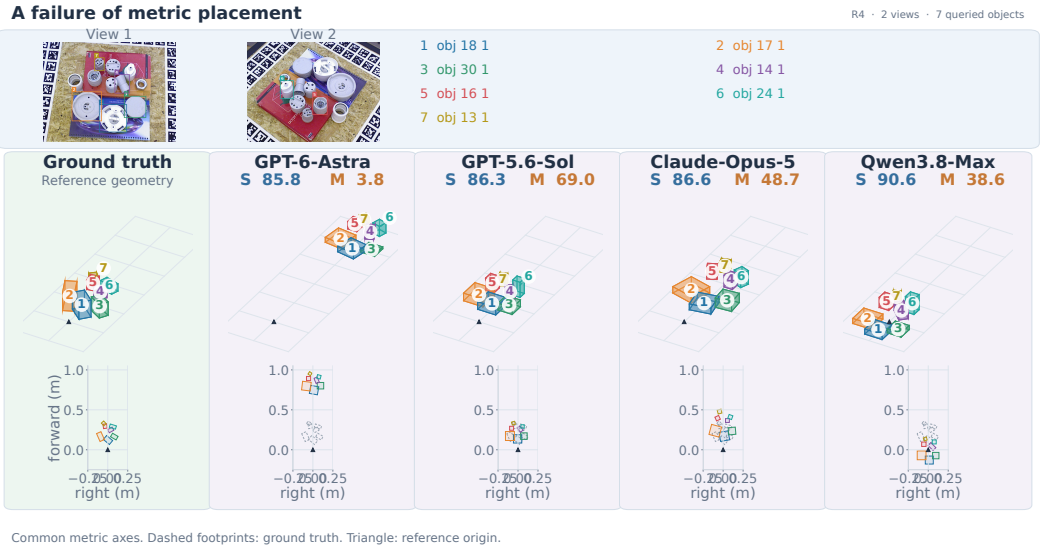


Figure 11: Camera-relative depth failure in tabletop case R4. GPT-6-Astra obtains high Structure but low Metric, despite returning all seven queried objects. The common axis limits expose the displacement of its object group away from the camera. GPT-5.6-Sol places the objects much closer to their GT positions. The contrast illustrates a scene-specific calibration failure of the aggregate leader.

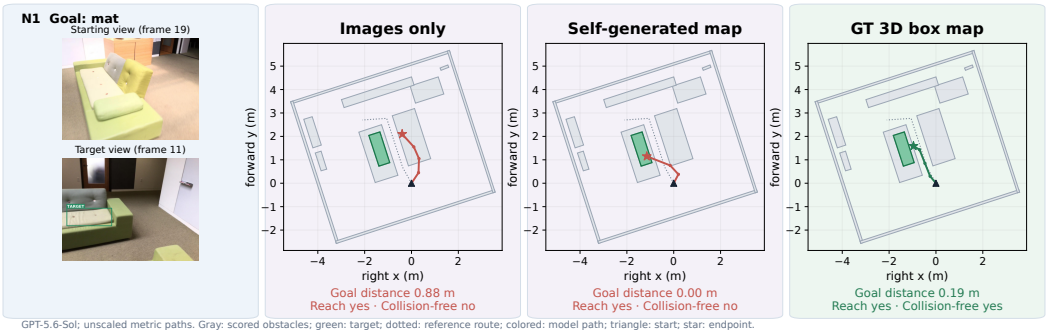


Figure 12: Navigation case N1 for GPT-5.6-Sol under the unscaled metric protocol, comparing image-only, self-generated 3D box map, and GT 3D box map conditions: goal proximity is insufficient when the path collides with obstacles. Gray polygons are the GT obstacles used by the scorer in every panel; green polygons are the target footprints. The goal object is excluded from collision checking. Dotted reference routes are illustrative; success does not require following that particular route.

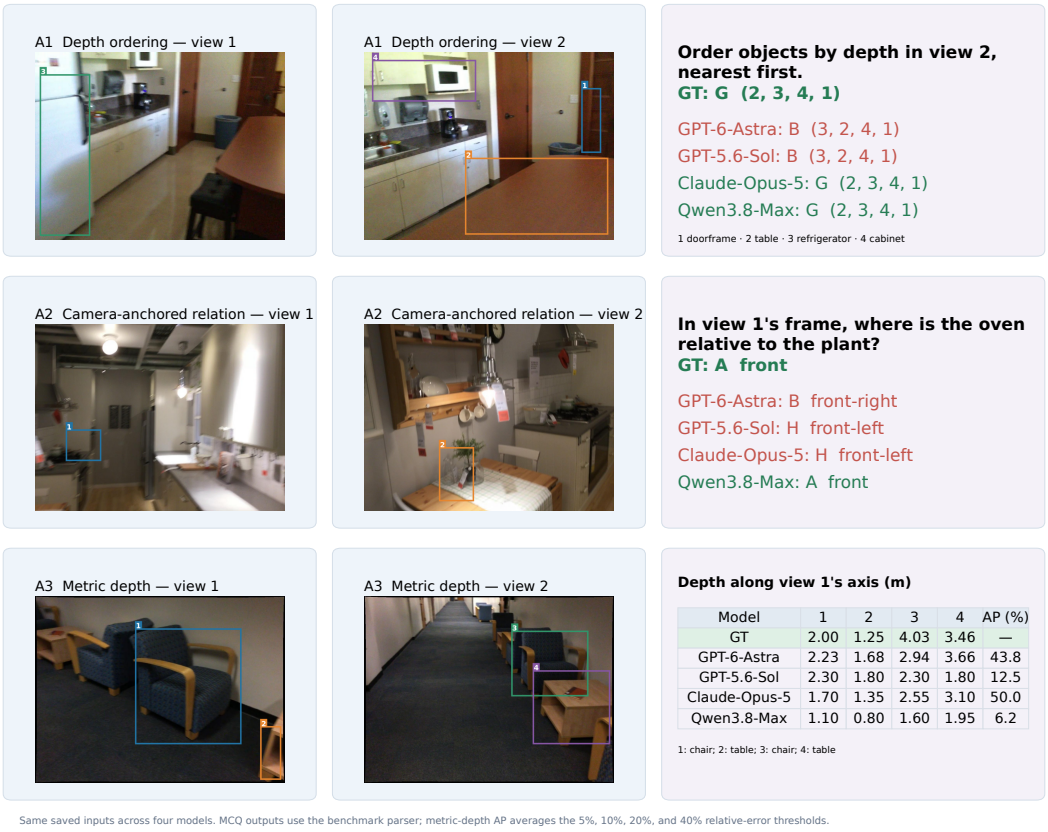


Figure 13: Three atomic cases with the same four models. Colored boxes and object numbers are overlaid from the original prompt annotations. A1 tests nearest-to-farthest order along view 2's axis; A2 tests an oven's direction relative to a plant in view 1's frame; A3 requests metric depths along view 1's axis, including objects marked in view 2. Green and red MCQ answers indicate correct and incorrect predictions. Metric-depth AP averages object accuracy at 5%, 10%, 20%, and 40% relative-error thresholds.